Clustering in efficient lexical design

immediate

April 24, 2014

Keywords: linguistics, lexical design, communication, phonotactics,

Abstract

1 Introduction

Saussure famously stated that there is an arbitrary relationship between word forms and their meanings. As Hockett (1960) wrote, "The word 'salt' is not salty nor granular; 'dog' is not 'canine'; 'whale' is a small word for a large object; 'microorganism' is the reverse." Our ability to manipulate such arbitrary symbolic representations is one of the hallmarks of human language and, in fact, makes language richly communicative since it permits reference to arbitrary entities, not just those that have iconic representations (Hockett, 1960).[1]

Exceptions to the arbitrariness of linguistic symbols are few and, in some sense, prove the rule. 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). 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). More broadly, it has been shown that speakers can distinguish between different classes of words, such as concrete versus abstract nouns, based on phonetic properties (Reilly et al., 2012). These examples are notable not for their regularity, but for their rarity and lack of systematicity. Even representations of animal noises, for instance, differ across languages (compare English "cockadoodledoo" to Spanish "cocorico"). Languages' choice of non-onomatopoeic words is even more the result of historical accident: there 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*. As a result of this widespread arbitrariness, there are many degrees of freedom in how a lexicon is structured. Lexicons can vary in their distribution of word lengths, in their distribution

[1] Monaghan et al. (2011) and Gasser (2004) note that arbitrariness may also be a boon to learning in that, if there were a tight relationship between form and meaning, semantically similar entities like *dog* and *cat* would have similar names (perhaps *dog* and *tog*). Such a system would increase the chance of perceptual confusability since those words are often used in similar contexts and so context alone would not always be able to disambiguate the words.

of phonemes, in how they carve out the semantic space, what wordforms they choose, and in many other properties.

The existence of an information storage and processing system that allows such a high degree of freedom and arbitrariness is unusual in human cognition. The way in which languages use this freedom is potentially informative about the underlying forces shaping human communication and thought. Specifically, we might expect that over time our capacity for arbitrary word meanings has led to an essentially random or non-systematic distribution of wordforms–at least among monomorphemic words, which are the focus of this study.[2] This landscape of randomly distributed wordforms is the default hypothesis a pure Saussurean might make. On the other hand, it is easy to imagine that in the face of minimal constraints beyond phonotactics on the set of well-formed words a language may use, lexicons might evolve under other pressures—perhaps for communicative or cognitive optimization.

In fact, communicative and cognitive optimization may yield very different types of lexicons. If we assume a noisy channel model of communication (Gibson et al., 2013; Levy, 2008; Shannon, 1948a), there is always some chance that the linguistic signal will be corrupted through errors in production, errors in comprehension, or some other form of noise in the signal. A lexicon is maximally robust to noise when the expected phonetic distance between words is maximized (Flemming, 2004; Graff, 2012), an idea used in error-correcting codes (Shannon, 1948b). Such sparsity has been observed in phonetic inventories (Flemming, 2002; Hockett & Voegelin, 1955) in a way that is sensitive to phonetic context (Steriade, 1997, 2001). Applying this idea to the set of wordforms chosen in a lexicon, one would expect wordforms to be maximally dissimilar from each other, given the bounds of phonotactics and conciseness.

The pressure for dispersion becomes apparent when one considers that it seems almost absurd to imagine a language that has many long words that sound almost identical to other long words. While English has words like ‘accordion’ and ‘encyclopedia,’ it certainly does not also have ‘accordiom’ or ’encyclofedia.’ Such forces may influence coining and borrowing; it seems plausible, for instance, that the 16th-century French borrowing ‘caparison’ (a type of horse covering) has remained very infrequent in English because it is extremely likely to be misheard as the far more frequent ‘comparison.’

Another force, however, conflicts with the pressure for sparsity in the lexicon: there is a pressure for the lexicon to be easy to learn, remember, produce, and process. In the most extreme possible case, one might imagine a language with only one wordform. Learning the entire lexicon would, in this case, be as simple as learning to remember and pronounce one word. While this example is extreme effects of systematicity and re-use can be found at every level of the lexcion. There are complex constraints that govern the set of sounds allowed in any given language (Hayes, 2012). Words can be regularly formed and stored using prefixes and suffixes (O’Donnell, 2011). Monaghan et al. (2011) show that this phonetic regularity helps language learners group words into categories. Storkel et al. (2006) show that adult learners more easily learn nonwords that appear in high-density phonological neighborhoods, where a phonological

[2] There is, of course, an obvious source of regularity in morphologically complex words like “happiness,” in which the words are formed from regular parts.

neighbor is defined as a word that differs from the target word by just one sound. These results are unsurprising since human memory is known to improve when information can be shared across different modalities.

The pressure for re-use manifests itself quite clearly in domains like phonology and morphology. While there are many possible sounds sequences that can exist in human languages, any given language relies on a relatively small subset of those sounds. And, by having regular suffixes like *ness* and *esque* that can derive new forms, it becomes possible to understand novel words through decomposition. It is this regularity that allows us to understand a word like *Twitter-esque*, even if we have never seen it before.

The basic challenge with assessing these forces—and comparing them to a truly arbitrary or non-systematic lexicon—-is that it is difficult to know what standard existing lexicons should be compared to. If we believe, for instance, that the wordforms chosen by English are sparse, we must be able to quantify sparseness *compared to* some baseline. Here, we present a novel method for solving this methodological puzzle, and we use it to study the fundamental forces shaping the set of monomorphemes chosen by the English lexicon[3].

2 Neighborhood and frequency effects

As a first pass at evaluating the extent to which languages show a preference for re-use, we investigated whether (a) words that are orthographically likely are more likely to be frequent than words that are less orthographically probable and (b) whether words in dense neighborhoods are more likely to be frequent than words with few neighbors.

2.1 Method

We used a corpus of 115 languages downloaded from Wikipedia and restricted our analysis to the top 20,000 most frequent wordforms in each language. We used exclusively orthographic wordforms here, which we believe are a reasonable proxy for phonetic forms. One natural test statistic is *phonological neighborhood density* (PND), which focuses on 1-edit pairs. PND is defined for each word as the number of other words in the lexicon that are one edit (an insertion, deletion, or substitution) away in phonetic space (Luce, 1986; Luce & Pisoni, 1998). For instance, ‘cat’ and ‘bat’ are phonological neighbors, as well as minimal pairs since they have the same number of letters and differ by 1. ‘Cat’ and ‘cast’ are neighbors but not minimal pairs. Neighborhood density has been shown to affect a wide variety of lexical processing from reaction time (Vitevitch & Luce, 1998) to fMRI activation of language areas (Prabhakaran et al., 2006).

For each word in each language, we computed the number of minimal pairs and neighbors that the word had among the 20,000 wordforms in the sample. We also

[3] Prior theoretical arguments about the statistics of language—in particular Zipf’s law Mandelbrot (n.d.); 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. This model makes extremely unrealistic assumptions about the true generative processes of language Howes (1968); Piantadosi et al. (Under review) and therefore provides a poor baseline for studying the true statistical forces of human languages.

trained a 3-phone model (smoothed and with backoff) on each set of 20,000 words and used the language model to find the probability of each word string under the model. EXAMPLE OF HIGH AND LOW.

2.2 Results

Figures 1 and 2 show density plots of scaled orthographic probability versus scaled log frequency and scaled number of neighbors versus scaled frequency, respectively, for 4-letter words. The positive slope of the red line reveals a correlation, in almost all languages, between these variables.

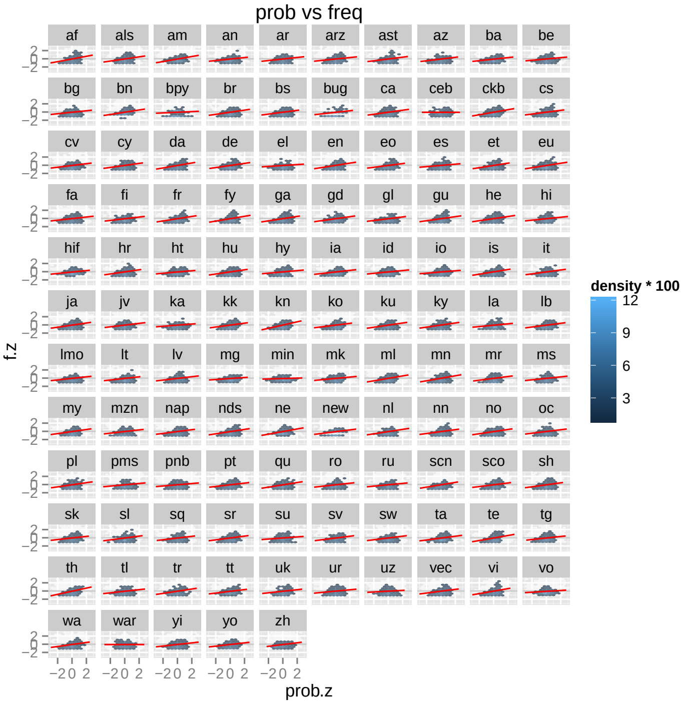

Figure 1: Orthographic probability plotted against frequency. The red line is a line of best fit.

Using Spearman correlations to obtain non-parametric correlation coefficients, analyzing each length separately, we found that almost all languages show significant correlations between log frequency and orthographic probability (mean r = .22 averaging across the individual correlation for each length and language), between log frequency and number of neighbors (mean r = .22), between log frequency and number of minimal pairs (mean r = .18), and (unsurprisingly) between orthographic probability and number of minimal pairs (mean r = .50) and between orthographic probability and number of neighbors (mean r = .54). Figure 3 shows this graphically by length.

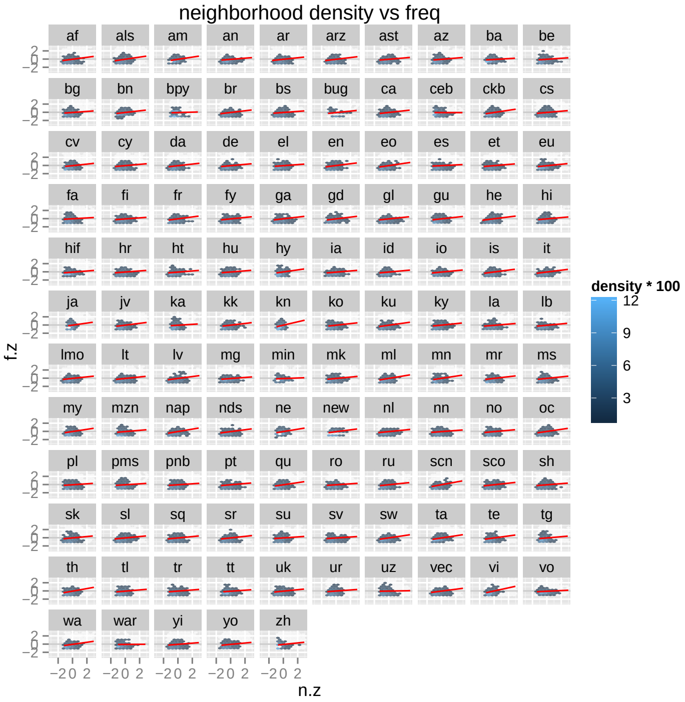

Figure 2: Number of orthographic neighbors plotted against frequency. The red line is a line of best fit.

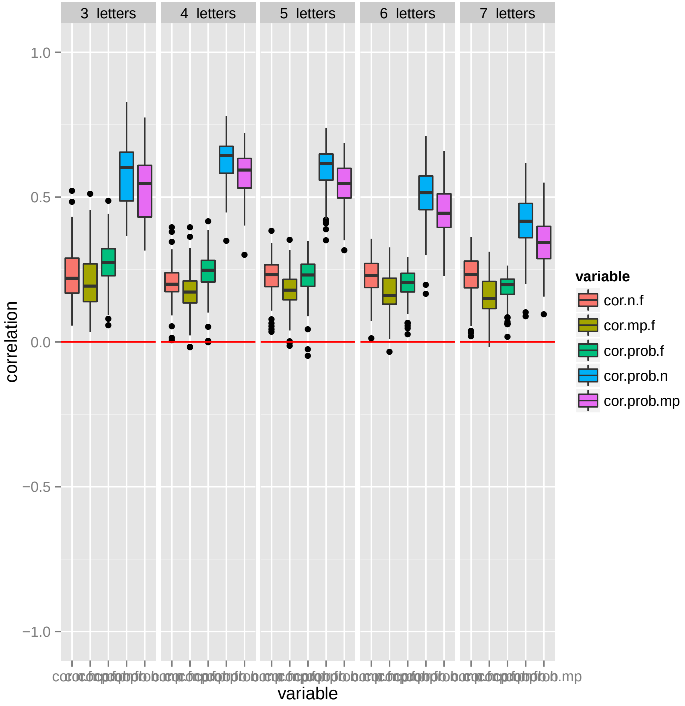

Figure 3: Each bar shows the mean correlation across languages for the variable shown. “n” here refers to number of neighbors, “f” is log frequency, “mp” is minimal pairs, and “prob” is orthographic probability as measured by the 3-phone model.

Table **??** shows the number of correlations for each variable, for each length, that are not significant out of 115. For each contrast, the vast majority of languages show a significant correlation in the expected direction.

	variable	3 letters	4 letters	5 letters	6 letters	7 letters
1	cor.mp.f.p	5	4	4	6	11
2	cor.n.f.p	3	3	2	2	4
3	cor.prob.f.p	3	3	3	2	2
4	cor.prob.mp.p	0	0	0	0	1
5	cor.prob.n.p	0	0	0	0	0

2.3 Discussion

Unsurprisingly, we found a robust correlation between orthographic probability and number of neighbors/minimal pairs. This result holds for all lengths across the vast majority of languages and is consistent with what we should expect. The word “set” is more likely to have neighbors in English than the word “quiz” simply because the letters in “set” are more common. So, probabilistically, there are more opportunities for a word to be orthographically close to “set” than to “quiz.”

Perhaps more interesting is the correlation between neighborhood density and frequency and between orthographic probability and frequency. A prior, it is not obvious that we should see this effect since frequency is not used in computing orthographic probability. The result suggests that, if a word has many neighbors and/or has high orthographic probability, it is more likely to be frequent.

One explanation for this relationship between frequency and orthographic probability is apparent when we imagine words being probabilistically generated by an orthographic model. That is, when we need a meaning for a new word, we generate a wordfrom from our 3-phone model. Words that are more likely will be independently generated more often, which would allow their frequency to increase. Consider homophones like “pore”, “pour”, and “poor’.” These words have no recent shared etymological ancestry but all ended up with the same phonetic pronunciation in English because the phonetic string is quite likely. Therefore, the frequency of the sound string is inflated by corresponding to multiple meanings. Thus, the phonetic probability could give rise to higher frequency for the string. And, of course, we already have seen that higher string probability gives rise to more neighbors.

In the next section, we present results from an experiment that compares the natural language lexicon to a plausible baseline.

CITE OLD RESULTS. BAAYEN?

3 Null lexicon experiments

3.1 Lexicons

Because we want to evaluate the real lexicons of mono-morphemic words relative to baselines, we focus on languages for which we could obtain reliably marked morphological categories. For English, German, and Dutch, we use CELEX pronunciations and restrict the lexicon to mono-morphemic words (words marked M). For French, we used Lexique, and I.D. (a native French speaker) identified mono-morphemic words by hand. For simplicity, we study as the 'real lexicon' a subset of the English lexicon consisting of all words of 4 to 8 phones[4] from the CMU corpus (Weide, 1998), as modified for the BLICK phonotactic probability calculator using the default English grammar with stress (Hayes, 2012).

In order to avoid the effect of homophones, we allowed only unique pronunciations. We define the lexicon to be the set of unique pronunciations in CELEX.

All three CELEX dictionaries were transformed to make diphthongs into 2-character strings. From the English CELEX, we removed a small set of words containing foreign characters. The lexicons used for each language are included in Appendix A.

3.2 Choosing the best null model

3.2.1 Method

In order to evaluate the real lexicon against a plausible baseline, we first defined a number of baseline lexical models, which we describe below:

- n-phone models: For n from 1 to 5, 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 letter after having just seen the previous *n-1* letters: $P(x_i|x_{i-(n-1)},...,x_{i-1})$. To avoid overfitting the model, we used backoff and smoothing. 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.
- n-syll models: For n from 1 to 2, we trained a language model over syllables. This generative procedure is like the n-phone model, but we instead train over syllables.
- PCFG: **ISABELLE**

In order to evaluate their ability of each model to capture the structure of the real lexicon, we trained each model on 75% of the lexicon and evaluated the probability of generating the remaining 25% of the lexicon.

[4] Three-phone words were excluded because they tend to have a phonotactics distinct from longer words (typically only CVC), and longer words were excluded because any two words of 9 letters or more are unlikely to have very much overlap, so it becomes difficult to compare.

3.2.2 Results

The parameter sweep revealed that the optimal smoothing parameter was .01. In all models described, unless otherwise stated, we use .01 as the smoothing parameter.

As shown in , the 5-phone model is the best performer across all languages.

PUT FIGURE HERE

3.3 Results

We simulated 30 lexicons of each type and computed test statistics for the real lexicon and each of the simulated lexicons. As described above, we then compared the distribution of this statistic across the simulated lexicons to its value in the real lexicon and compared the real lexicon to each of the null models. Below, we describe these measures and use them to characterize how the simulated lexicons from the generative model compare to the real lexicon of each language that we tested.

To compare 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. Just like in classic null hypothesis testing, we compute a z-score using the mean and standard deviation estimated from the 50 lexicons generated by our "null" 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 in the tables reflects the probability that the real lexicon value could have arisen by chance under the most tightly constrained generative model.

3.3.1 Minimal pairs

Figure 4 summarize these hypothesis tests, showing how the various simulated lexicons compare to the real lexicon in terms of log number of minimal pairs per 10,000 words for words of varying length. For words of 6 phones and above, minimal pairs are quite rare, but the same pattern holds. Each color represents a different type of generative model. The red dot represents the real lexicon value.

Because not all sounds are equally confusable ('p' and 'b' are far more confusable than 'p' and 'e'; see Graff (2012)), we also look specifically at phonological neighbors whose minimal difference relies on a subtle contrast between a voiced and unvoiced consonant. To check whether the clumpiness effect holds even for minimal pairs that are phonetically confusable (i.e., minimal pairs that differ only by a voicing contrast as in pig/big and mad/mat), we additionally compared the real lexicon and the simulated ones in terms of minimal pairs that differ only in phonetically similar stop consonants: 'p/b', 't/d', and 'k/g'. The real lexicon has more such minimal pairs than even the most constrained generative model. minimal pairs from the most tightly constrained generative model (length, CV, and phonotactically matched). The red dot represents the real lexicon value, and the dotted lines represent 95% confidence intervals based on the simulated values. In all cases, the red dot falls well to the right of the dotted line, indicating that the real lexicon has significantly more of each type of minimal pair.

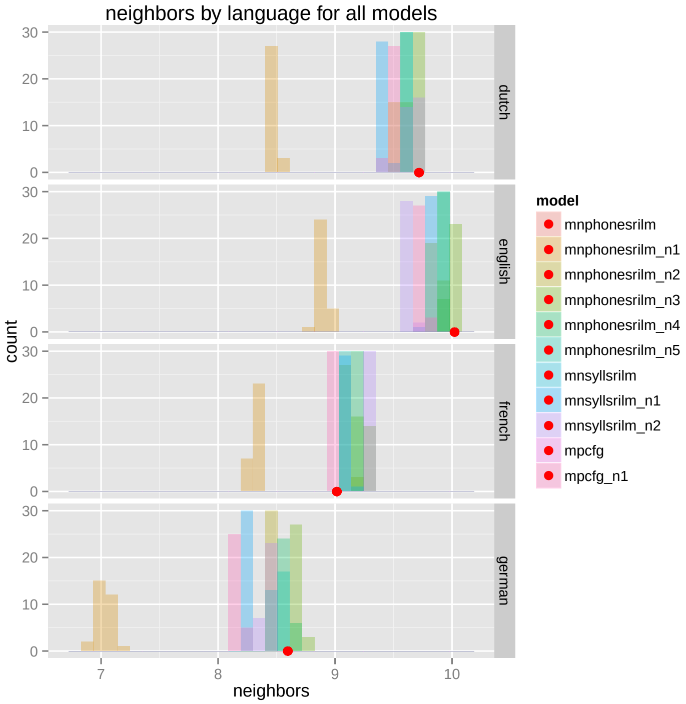

Figure 4: Each colored histogram represents a distribution of minimal pair counts, broken up by word length in phonemes, across the 50 simulated lexicons from each of the four generative models. The y-axis shows the log number of minimal pairs per 10,000 words for each of the four types of lexicons. The red dot shows the real lexicon value for each condition. The most tightly constrained generative model (CV and phonotactically matched) comes closest to matching the real lexicon value in all cases, but all of the histograms for all of the lengths fall to the left of the red dot, which suggests that the real lexicon has more minimal pairs than any of the simulated ones.

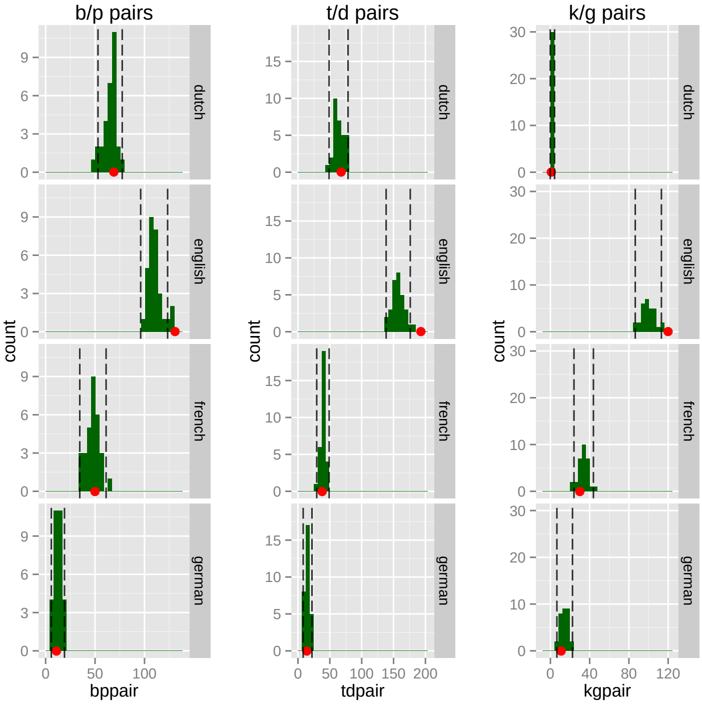

Figure 5: These histograms show the distribution of the most closely matched simulated lexicon (in green) compared to the real lexicon (the red dot) in terms of confusable minimal pairs: 'p/b', 'k/g', and 't/d'. The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. In all cases, the red dot is significantly to the right of the simulated lexicon distribution– which suggests that the real lexicon is clumpier than expected by chance.

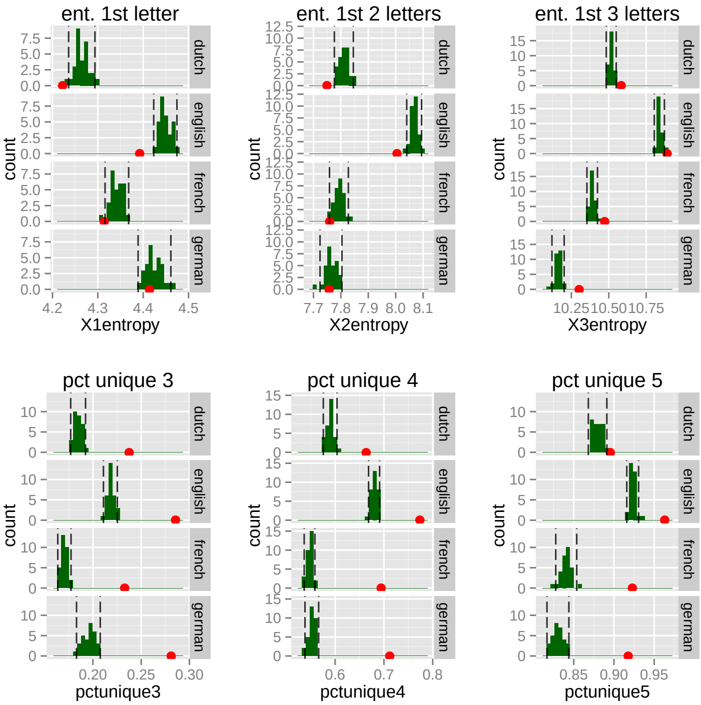

Figure 6: NEED

3.3.2 Word onset measures

Besides phonological neighbors, there is evidence that word onsets are of particular importance in lexical processing (Marslen-Wilson, 1980, 1987; Wingfield et al., 1997), so we also compute several test statistics that focus on shared word onsets. For instance, we looked at the percent of the lexicon that could be disambiguated after seeing only *n* sounds.

To see whether word onsets show clustering in the real lexicon, we looked at the percent of words in the real lexicon that are able to be uniquely disambiguated after 3 sounds and 4 sounds. The words *cap* and *cat*, for instance, can be disambiguated from each other after 3 sounds, but not Figure 6 shows histograms for each of these measures. The red dot representing the real lexicon shows that the real lexicon has significantly more confusable onsets than the simulated lexicons.

3.3.3 Levenshtein distance

We can evaluate clustering using more global measures by considering the average string edit distance (*Levenshtein distance*) between words. The Levenshtein distance between two sound strings is simply the number of edits required to get from one string

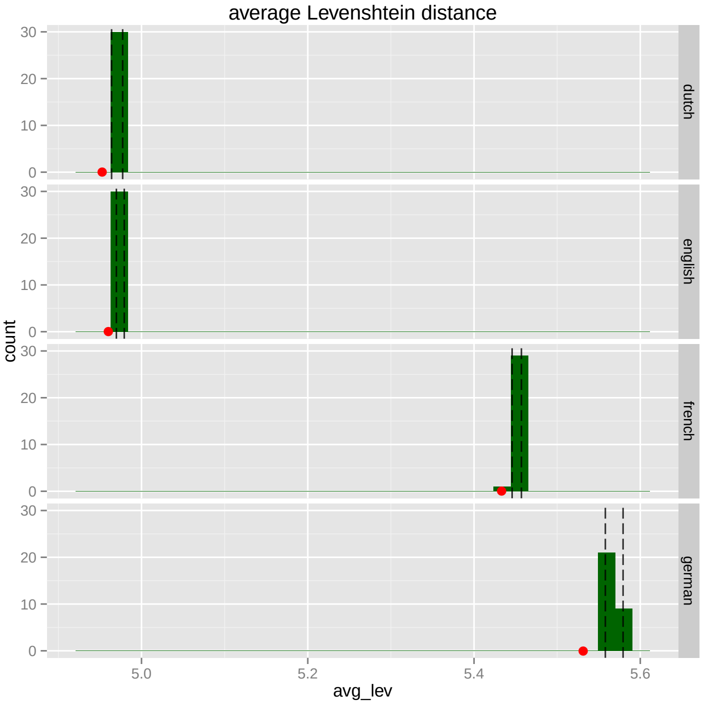

Figure 7: The histograms show the distribution of average Levenshtein distances for each of the 50 simulated lexicons (restricted to only the most closely matched generative model). The red dot represents the real lexicon's value, and the dotted lines are 95% confidence intervals.

to another.[5].

In the remainder of the results, we will focus mainly on the most tightly generative model (CV and phonotactically matched)

3.3.4 Network measures

Simply calculating phonological neighbors and onset confusability, however, does not tell us everything about how sounds are distributed across a lexicon. Do certain words have many neighbors and others very few? Or do neighbor pairs tend to be more uniformly distributed across the lexicon? To answer these questions, we construct a phonological neighborhood network as in Arbesman et al. (2010), whereby we build a

[5] So, the Levenshtein distance between 'cat' and 'cast' is 1 (insert an 's'), and it is 2 between 'cat' and 'bag' (c → b, t → g). 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.

graph in which each word is a node and any phonological neighbors are connected by an edge, as in the toy example in Figure 8.

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 better characterize the clustering behavior of the lexicon. We use measures like *average clustering coefficient*, *transitivity*, and percent of nodes in the *giant component* to evaluate how tightly clustered nodes in a network 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). Thus, transitivity in effect asks, given A is connected to B and B is connected to C, how likely is it that A is also connected to C? The average clustering coefficient is a closely related measure that find the average clustering coefficient across all nodes, where the clustering coefficient of a node is defined as the fraction of possible triangles that *could* go through that node that actually do go through that node. Both of these 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. Applying these measures to the lexicon gives insight into lexical structure not captured by just the number of neighbors or by Levenshtein distance.

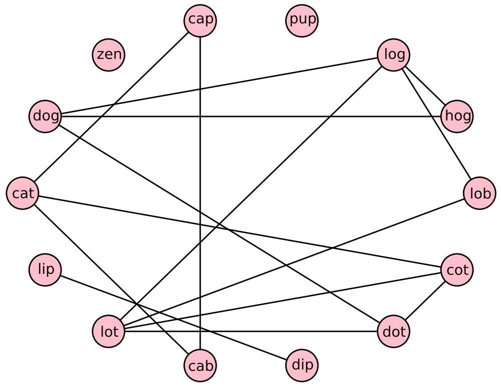

Figure 8: Example phonological network. Each word is a node, and any words that are 1 edit apart are connected by an edge.

Figure 9 shows examples of such networks, where each word is a node, with an edge drawn between any two words that are phonological neighbors (1 edit away).

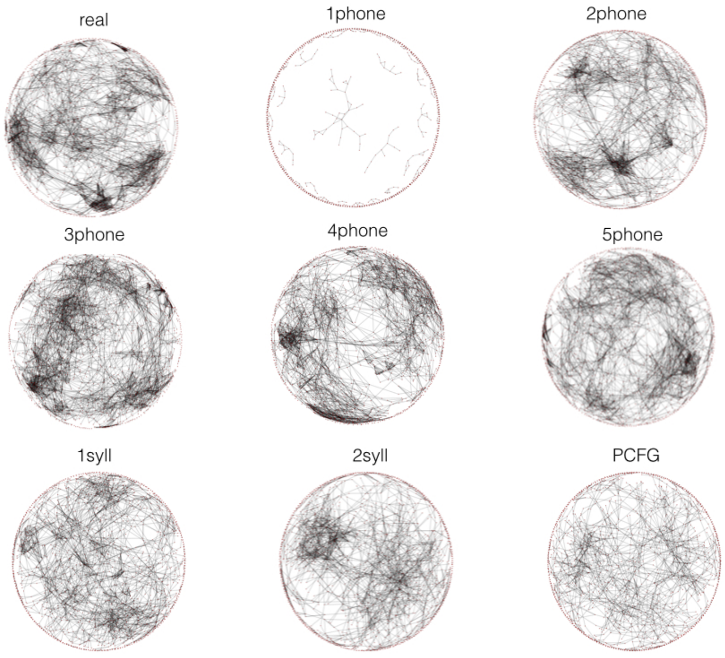

Figure 9: Sampling of phonological neighbor network from the different models. Each point is a word, and any two connected words are phonological neighbors. The simulated lexicons from less constrained generative models are less clustered and have more isolates (words with no neighbors, plotted on the outside ring).

Words with no or few neighbors are clustered on the outside. Words with many neighbors are plotted more centrally. As can be seen in Figure 9, substantially more clustering is observed in the more restrictive generative models, but the real lexicon is, overall, significantly clumpier than the lexicons produced by any of the generative models.

4 Effect of semantics

As discussed above, there are certain non-trivial effects of semantics in the structure of the lexicon, such as the presence of semantic clusters like *gl-* words. It is also, of course, the case that certain words are etymologically related or derived from other words in the lexicon (even when the lexicon is restricted to morphologically simple words). For example, *skirt* and *shirt* are historically the Old Norse and Old English form of the same word, whose meanings have since diverged. Both sources of structure could potentially contribute to pockets of words that are both phonologically and semantically similar to each other.

To assess whether the presence of these sorts of clusters is driving the clumpiness in the lexicon, we additionally analyzed the lexicon controlling for semantic factors by

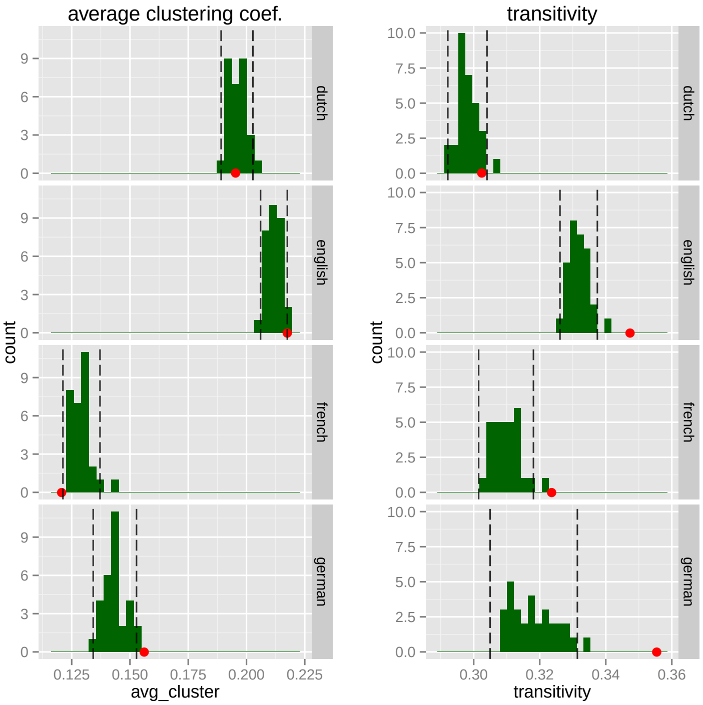

Figure 10: These histograms show the distribution of the most closely matched simulated lexicon (in green) compared to the real lexicon (the red dot) in terms of network measures for lexical networks (where each node is a word and any 2 nodes that are minimal pairs are joined in the network): the percent of nodes in the giant component, the average clustering coefficient, and transitivity. In all cases, the red dot is significantly to the right of the simulated lexicon distribution–which suggests that the real lexicon is clumpier than expected by chance.

defining semantic relatedness scores for each pair of words in the real lexicon and in the simulated lexicons. Using Fellbaum (2010), we calculated the *minimum* semantic path distance between any two words of the same primary part of speech (as defined by CELEX) in the real lexicon[6]. This gives us a semantic relatedness score between any two words in the real lexicon.

To compare the correlation between semantic relatedness and phonological relatedness in the real lexicon, we need to compare it to a baseline. One idea would be to randomly permute the semantic relatedness scores from the real lexicon and arbitrarily assign them to the pairs in the simulated lexicons. But this glosses over an important source of regularity. It turns out that word pairs that have fewer semantic neighbors systematically tend to have lower phonotactic probabilities and fewer phonological neighbors. Therefore, word pairs consisting of low-probability words should also receive lower semantic relatedness scores. Therefore, we matched each word in the simulated lexicon to a word in the real lexicon by sampling from the appropriate length, CV, and phonotactic probabiltiy bin. That is, in one lexicon, *jingle* was matched to *chardle* and *jangle* to *vorset* (orthographic instead of phonetic transcriptions used for clarity). We assigned the real pair's semantic similarity score to their simulated matched bin. Because the semantic path distance between *jingle* and *jangle* is 1 (the maximum similarity score), the 'simulated' semantic distance between *chardle* and *vorset* is also 1.

For each noun in the real lexicon (items listed exclusively as nouns in CELEX), we found its Levenshtein distance to every other noun. Each of these noun pairs has a matched pair in the simulated lexicon (as with jingle/jangle and chardle/vorset). We also found the Levenshtein distance for each of the matched pairs in the simulated lexicon. We can then look at the difference between the real and simulated Levenshtein distance ($RealPairLevenshteinDistance - SimulatedPairLevenshteinDistance$) and the effect of semantic similarity on phonological similarity.

Specifically, we examined the effect of Wordnet path similarity on Levenshtein distance in a representative lexicon. Because Wordnet is not reliable for comparing semantic similarity across part of speech categories, we focused on nouns and restricted the analysis to 4-phone words. Figure **??** shows the average difference between the real lexicon Levenshtein distance and the matched pair Levenshtein distance (from a representative CV and phonotactically matched generative model) on the y-axis. The x-axis is Wordnet semantic similarity. Closely semantically related words are likely to have a smaller Levenshtein distance than two

One class of words that one might expect should be particularly phonetically distinct is antonyms, like *hot* and *cold*. Words in such pairs can typically occur in the same linguistic context but have critically different meanings. In a study of antonyms pairs drawn from Wordnet and restricted to be mono-morphemic using CELEX, we found that a word is on average more phonetically similar (using Levenshtein distance) to its antonym than to a random word that is not its antonym. For 43 antonym adjective pairs (86 unique words), the mean Levenshtein distance between a word and its antonym was 3.70 compared to a mean distance of 4.10 between a word and one of the remaining 84 words ($t = 2.26$, $p < .05$).

[6] That is, for any two words in a pair, if their primary meanings were quite different but their second or third meanings were very similar, the alternate meanings were used for determining the similarity measure.

5 Discussion

We have shown that the English lexicon uses its degrees of freedom in a systematic and interesting way. While we can still characterize the relationship between word forms and meanings as arbitrary, structure emerges when one considers the relationships within the space of possible wordforms. The real lexicon is clumpier than expected by chance, even when controlling for semantic relatedness of words. Across a wide varitey of measures of phonological confusability and similarity, the real lexicon shows significantly more clustering than even lexicons produced by the most tightly matched generative model.

Because we focused on mono-morphemic words, the 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 or by a state-of-the-art phonotactic model. Certainly, 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 null models. But, since even words that are very distantly related are more phonetically similar than their matched pair in the simulated lexicon, the lexicon's clumpiness cannot be attributed only to these factors. Rather, the clumpiness reveals a fundamental drive for regularity in the English lexicon, a drive that conflicts with the pressure for words to be as phonetically distinct as possible. This interplay between the competing pressures for sparsity and re-use could underlie not just the structure of the lexicon discovered here but also the pattern of word learning results shown in Storkel et al. (2006).

One possible source of the lexicon's clumpiness is that speaker preferentially re-use common articulatory sequences. That is, beyond just phonotactics and physical constraints, speakers 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 is "overfit": any new word is deeply dependent on the existing words in the lexicon.

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. The fact that more semantically related words are also more phonetically related supports this hypothesis. 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.

The interaction of these cognitive and articulatory constraints with the pressure for communicative efficiency is complex. 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. One possibility is that context is usually enough to disambiguate words, and 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 the dozens meanings for short words like *run*, does not impede communication and in fact abets 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.

More broadly, the methodology used here, whereby the real lexicon is compared to a distribution of statistically plausible 'null' lexicons, could be generalized to answer other questions about the the lexicon and human language more generally. While much previous work has focused on simply measuring statistical properties of natural language, modern computing power makes it trivially easy to simulate thousands of different languages with different constraints, structures, and biases. By comparing real natural language to a range of simulated possiblities, it is possible to assess what aspects of natural language occur by chance and which exist for a reason. We hypothesize that other languages will pattern similarly to English, but most languages lack the tools necessary to adequately calculate phonotactic probability for the analysis presented here. In future work, we hope to address other languages in order to see whether the clumpiness of the lexicon varies based on other properties of the language, such as its propensity for borrowing words from other languages.

In future work, it may be possible to test increasingly sophisticated models of phonotactics using this methodology. Perhaps 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 of natural language that can approximate natural language will need to account for not just 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 just on an underlying model of English sound structure but on the fact that "lax" is a word, that "zax" is not, and on countless other properties of the existing lexicon.

Indeed, we have shown that the English lexicon is more richly structured than previously thought. The space of English words is clumpier than any of the null models by a wide variety of measures: minimal pairs both across and within confusable sets of sounds, network properties, uniqueness of word onsets, and (for shorter words) edit distance. This property of the lexicon cannot currently be explained by any existing phonotactic models of English, and we thus conclude that underlying the pressure for sparsity in the lexical system is a deep drive for regularity and re-use beyond standard levels of lexical and morphological analysis

References

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010, March). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, *20*(03), 679–685. Retrieved 2013-02-07, from http://www.worldscientific.com/doi/abs/10.1142/S021812741002596X doi: 10.1142/S021812741002596X

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].*

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

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311. Retrieved 2013-04-16, from `http://muse.jhu.edu/content/crossref/journals/language/v080/80.2bergen.pdf` doi: 10.1353/lan.2004.0056

Fellbaum, C. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*, 231âĂŞ243.

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.

Gasser, M. (2004). The origins of arbitrariness in language. In *Proceedings of the annual conference of the cognitive science society* (Vol. 26, pp. 4–7).

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*. Retrieved from `http://www.pnas.org/content/early/2013/05/01/1216438110.abstract` doi: 10.1073/pnas.1216438110

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

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

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., & 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.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126âĂŞ1177.

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.

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

Marslen-Wilson, W. D. (1980). Speech understanding as a psychological process. In *Spoken language generation and understanding* (pp. 39–67). Springer.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102.

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

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. Retrieved 2013-04-04, from `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0022924` doi: 10.1037/a0022924

O'Donnell, T. (2011). *Productivity and reuse in language*. Unpublished doctoral dissertation, Harvard University.

Piantadosi, S., Tily, H., & Gibson, E. (2012, March). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. Retrieved 2013-04-02, from `http://linkinghub.elsevier.com/retrieve/pii/S0010027711002496` doi: 10.1016/j.cognition.2011.10.004

Piantadosi, S., Tily, H., & Gibson, E. (Under review). Information content versus word length in natural language: A reply to ferrer-i-cancho and moscoso del prado martin.

Prabhakaran, R., Blumstein, S. E., Myers, E. B., Hutchison, E., & Britton, B. (2006). An event-related fmri investigation of phonological-lexical competition. *Neuropsychologia*, *44*(12), 2209–2221.

Reilly, J., Westbury, C., Kean, J., & Peelle, J. E. (2012, August). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PLoS ONE*, *7*(8), e42286. Retrieved 2012-11-29, from `http://dx.plos.org/10.1371/journal.pone.0042286` doi: 10.1371/journal.pone.0042286

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

Shannon, C. (1948a). A mathematical theory of communication. *Bell System 1266 Technical Journal*, *27*, 623-656.

Shannon, C. (1948b). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

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., Armbruster, J., & Hogan, T. P. (2006, December). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. Retrieved 2012-11-29, from `http://jslhr.asha.org/cgi/doi/10.1044/1092-4388(2006/085)` doi: 10.1044/1092-4388(2006/085)

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

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.

Weide, R. (1998). *The CMU pronunciation dictionary, release 0.6.* Carnegie Mellon University.

Wingfield, A., Goodglass, H., & Lindfield, K. C. (1997). Word recognition from acoustic onsets and acoustic offsets: Effects of cohort size and syllabic stress. *Applied Psycholinguistics*, *18*, 85–100.