Finding structure in the lexicon

-Saussure famously stated that there is an arbitrary relationship between the signifier and signified, but there is of course much structure in the lexicon word length (Piantadosi et al., 2011), phonetic dispersion (Fleming 2004, Graff 2012), and sound-symbolic relationships between semantics and phonetics. In the extreme case, we can observe that no known language has exclusively very long words or only one consonant or only words that differ by one phoneme from all other words in the lexicon. Is this a result we should expect by chance, or does it reveal a deeper property of natural language design?

-In order to investigate structure in the lexicon below the level of the morpheme, it is necessary to develop a model of what a plausible baseline lexicon looks like. Mandelbrot and others have sometimes used a random-typing model (the “monkey model”) as a baseline for comparison, but this model is radically unlike the generative process that underlies word formation in natural language.

-Here, we propose several models for generating “null” English lexica that can then be compared against the real English lexicon. In our real lexica, words are sampled i.i.d with varying levels of constraint. By comparing the simulated lexica to the real lexicon, we can then ask how the presence of other words in the lexical network affects the probability of finding word W in lexicon L.

-If the real lexicon is sparser than expected by chance, that would suggest a drive towards more easily distinguishable forms. If the real lexicon is clumpier than expected by chance, that would suggest a preference for the preferential re-use of sounds in new words.

The lexicon is phonologically clumpy

Lexical networks

Minimal pairs in 4 and 5 phone lexicons

Results

Compared to the null lexicon, the real lexicon is clumpy by numerous measures of clustering.

- There are both more minimal pairs and more 2-edit pairs in the real lexicon than in any simulated lexicon. This holds for specific lengths as well as across the entire lexicon.

- Network measures for lexical neighbor networks, (the giant component percentage, transitivity, average clustering) reveal that the real lexicon is more tightly clustered.

- Word beginnings are less unique in the real lexicon, i.e., it takes longer to disambiguate a word on average.

- The effect holds even for minimal pairs known to be confusable (voiced/unvoiced pairs).

Simulated lexica vs. real lexicon

Effect of semantic clustering

-A difference between the real lexicon and the simulated ones is the presence of semantic groups. We find that nouns are more similar to nouns, and verbs are more similar to verbs than either is to the other. Antonyms are more similar to each other in Levinshnein distance than they are to other words.

-To assess semantic effects on phonological clustering, we used Wordnet to measure semantic path distance for each possible part-of-speech matched word pair in the lexicon. We divided the pairs into semantically related (top 25% of pairs in semantic relatedness) and semantically unrelated (bottom 25%). The most semantically related pairs of words in the lexicon are likely to have a smaller Levenshtein distance and are more likely to be minimal pairs than a pair of words that is semantically unrelated. The strong clustering among words that are 1 jump away in Wordnet can likely be attributed to sound symbolism and the presence of etymologically related word pairs (single/jangle).

- Even among the most semantically unrelated pairs in the lexicon, a pair of words in the real lexicon is more likely to have a small Levenshtein distance and is more likely to be a minimal pair than a word in the most constrained simulated lexicon.

Reasons for clumpiness

We propose and give evidence for several mechanisms by which lexical clumpiness arises:

sound symbolism: As shown above, some lexical clumpiness can be attributed to the effect of semantics. Reilly et al. (2012) find that subjects can classify abstract and concrete words based on phonetic properties, and Abelin (1999) finds psychological effects of phonosematology. These types of sound-symbolic effects, which exist in the real lexicon but not in simulated lexica, could give rise to some phonological clustering, but they are not the only source since the effect holds even for semantically distant words.

learning bias: Storkel et al. (2006) find that adults more easily learn words in high-density neighborhoods than those in low-density neighborhoods. This preference in learning could contribute to high-neighborhood learning bias.

preferential re-use of sound sequences: Given that speakers infer phonotactic constraints from the lexicon, the constraints that they learn will be inherently biased towards the words that already exist in the lexicon.

Conclusions

-Across a wide variety of metrics, the real lexicon is clumpier than expected by chance. The effect is stronger among semantically related words but holds even for semantically distant words.

- We believe this reveals a fundamental structure inherent in natural language below the level of the morpheme.

Citations


-Saussure famously stated that there is an arbitrary relationship between the signifier and signified, but there is of course much structure in the lexicon word length (Piantadosi et al., 2011), phonetic dispersion (Fleming 2004, Graff 2012), and sound-symbolic relationships between semantics and phonetics. In the extreme case, we can observe that no known language has exclusively very long words or only one consonant or only words that differ by one phoneme from all other words in the lexicon. Is this a result we should expect by chance, or does it reveal a deeper property of natural language design?

-Simulating “null” lexica

-To simulated null lexica, we trained a 6-phone model on the real lexicon, where the real lexicon is taken from Hayes CMU corpus for Blick phonotactic probability calculator (Hayes 2012) and restricted to mono-morphemic words that appear in CELEX. Candidate words are generated from the model.

-All simulated lexica are sampled to match the real lexicon for length distribution (same number of 4- phone words, 5-phone words, etc.) and are restricted to 4 to 8 phones.

-For the most restrictive simulation, we scored all candidate words on the Blick phonotactic probability calculator and sample to match the real lexicon in distribution of phonotactic probability and CV pattern.

<table>
<thead>
<tr>
<th>lexicon type</th>
<th>generative process</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>real lexicon</td>
<td>from Hayes CMU corpus for Blick and further restricted to CELEX monomorphemic words (i.e., words categorized as M) of 4 to 8 letters (n = 18,279)</td>
<td>brick</td>
</tr>
<tr>
<td>random n-grams</td>
<td>10,000,000 n-phones trained on real lexicon and randomly sampled to match real lexicon for length</td>
<td>lara</td>
</tr>
<tr>
<td>Blick-matched lexicon</td>
<td>matched to real lexicon for length and phonotactic probability distribution as measured by Blick</td>
<td>dholi</td>
</tr>
<tr>
<td>CV-matched lexicon</td>
<td>matched to real lexicon for length and CV pattern</td>
<td>blott</td>
</tr>
<tr>
<td>Blick-matched CV-matched lexicon</td>
<td>matched to real lexicon for length, CV pattern, and phonotactic probability as measured by Blick</td>
<td>drock</td>
</tr>
</tbody>
</table>

Acknowledgments

We thank the CUNY 2013 reviewers, Alexa Khan, Peter Graff, Richard Futrell, and other members of Tedlab for helpful comments.

Kyle's contact info: kylemaho@mit.edu