Uncommon sounds for common words: Balancing phonemic type and token frequency within words for a higher entropy lexicon

Adam King*[1], Andrew Wedel[2]

*Corresponding Author: adam.king.phd@gmail.com
[1]GumGum, Santa Monica, USA
[2]Department of Linguistics, University of Arizona, Tucson, USA

It is a common empirical finding that the token frequency of a phonological contrast is positively correlated with its type frequency; sounds that occur most frequently across a corpus tend to appear in more unique word types and vice versa. This would naturally follow if contrasts are randomly distributed across words, though, given that sounds are used to distinguish words, this is precisely the opposite of what we expect. Lexical access proceeds incrementally, such that each successive sound gives a listener the required information to exclude a growing set of incompatible words as the speech stream is perceived (van Son & Pols 2003; Magnuson et al. 2007). If language is shaped to be an efficient system of communication, it would be expected that words in the lexicon would share similarities with a Huffman code, where each contrast equally divides the remaining words into probabilistically equal groups, causing instead a negative relation between the token frequency of a sound and the number of words it appears in at each branch-point, creating a more balanced contrast (Fig 1).

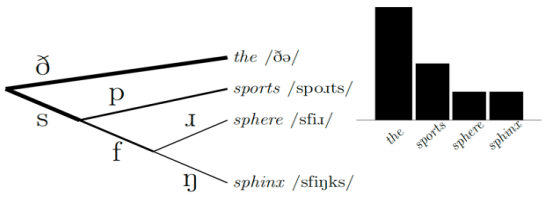

Figure 1. Contrast structure of a toy version of English. The bar chart on the right represents the probability of each word. On the left is a branching structure diagram of word contrasts where line thickness represents summed word probabilities along that branch. Each successive branch divides the remaining potential words into probabilistically balanced groups.

Here, we provide evidence for this predicted inverse relationship in a genetically balanced set of 20 languages, comparing the type and token frequencies of word-

initial biphones. First, we show that the commonly-noted positive type/token correlation is an artifact of the floor formed by the least frequent words in a corpus: if a word occurs only once, the sounds in it must occur at least once; if a word occurs twice, the sounds within it must appear twice, etc. When biphones that make up this floor are removed, that is, word-initial biphones that together comprise less than 5%[1] of all tokens in the corpus, we find instead a strong *negative* correlation between biphone type frequency and the mean token frequency of the words in which the biphone is found (Fig 2), leading to a more balanced contrast than would be expected if efficient phonological information transmission were not a shaping force on the lexicon.

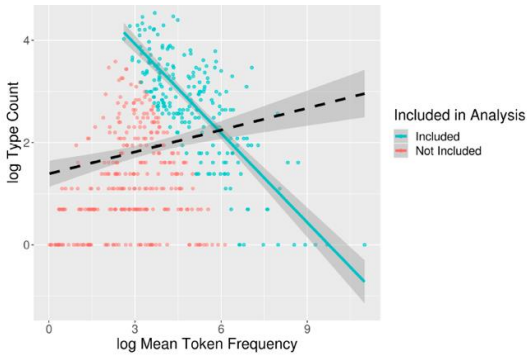

Figure 2. Relationship between mean token frequency and type count for word-initial biphones, excluding the least frequent 5% in the lexicon. For the top 95%, sounds that appear in fewer words tend to appear in higher frequency words overall, creating a more balanced contrast.

How might this balance in lexicons arise throughout language change? A large body of evidence shows that high-information segments tend to be hyperarticulated, while lower-information segments tend to be reduced (e.g., van Son & Pols 2003, Wedel et al. 2018), causing less frequent words to retain complex, marked segments over time, while more frequent words do the opposite. As less frequent words make up the large majority of a language's unique word types, these patterns of change should result in a lexicon in which less frequent words are composed of a wider diversity of sounds and sound sequences, as found in King & Wedel (2020). As a result, at a contrast point, sounds that are more often found in less frequent words should lead to relatively larger remaining cohorts, resulting in the expected negative type/token correlation. This represents a plausible mechanistic pathway from speaker-level micro-effects of information transmission toward a lexicon that is structured for higher entropy.

[1] The pattern remains the same for different threshold, e.g., 1%, 10%. We use 5% here for visualization purposes.

References

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

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*, 133-156.

Van Son, R. J. J. H. & Pols, Louis C. W. (2003). How efficient is speech. *Proceedings of the institute of phonetic sciences, 25, 171-184. University of Amsterdam*.

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language, 100*, 61-88.