# THE LEXICON AS A HUFFMAN CODE: WORDS ARE STRUCTURED FOR PROBABILISTICALLY BALANCED CONTRASTS

Adam King*[1] and Andrew Wedel[1]

*Corresponding Author: adamking@email.arizona.edu
[1]Department of Linguistics, University of Arizona, USA

## 1. Background

The key requirement for an efficient communication system is the maximization of communicated information across the message, relative to its length (Shannon, 1948). For simplicity, words can be thought of as messages, meaning that for efficient linguistic communication, word-internal information should be densely encoded within sub-word units.

As it happens, the lexicons (roughly, the set of words in the language) of nearly all tested languages display a robust relationship between word probability and length, i.e., *Zipf's law of abbreviation* (Zipf, 1935; Piantadosi, Tily, & Gibson, 2011; Bentz & Cancho, 2016), suggesting that the pressure for efficient communication partially affects the shape of words, at least with respect to the number of sub-word units. However, while short words may be part of an efficient system, they do not entail efficiency per se. For example, a language where every word is /ba/ satisfies the pressure for short forms, but fails *a priori* as a communication system. To this end, if the lexicon is structured for efficient communication, the sub-word units that build up full word forms should also be structured to be maximally informative, in addition to a general pressure for short lengths.

By definition, the average information for part of a message (in this case, contrastive sub-word units) is maximal when each possible contrast is equiprobable, such as in a Huffman code (MacKay, 2003). Therefore, if the lexicon is structured to be an efficient communicative code, contrasts within the lexicon's words should be between a more or less probabilistically balanced set of phonological contrasts.

## 2. Methodology

In this talk, we will show that contrasts in the lexicon significantly demonstrate effects of balance between contrastive phonological material, using a

dataset of phonetically transcribed lexicons of 25 typologically diverse languages. Furthermore, we will demonstrate that the predicted trend towards balance extends beyond what might be expected in any language-like code, by comparing to a baseline for each language (Prado Martin, 2013).

Primarily, we will show an inverse relationship between type and token frequencies for 95% of the most probable contrasts[1], a relationship expected of a probabilistically balanced distribution, and importantly, one that does not require a prior specification of "*how balanced*" it need be. Assuming that the probability for each contrast is more or less equal, those that are associated with fewer word types should be associated words that have a higher frequency, on average (Fig. 1).

As an example, consider the segment [ð] in English which begins few words but, of those it begins, many words are high frequency function words, e.g., *this*, *the*, etc. [t], on the other hand, begins many more word types, though the average frequency of [t]-initial words is less. When put together, the overall probability of a word token beginning with [ð] or [t] is more or less balanced, at least more so than if the relation between type and token frequencies was different.

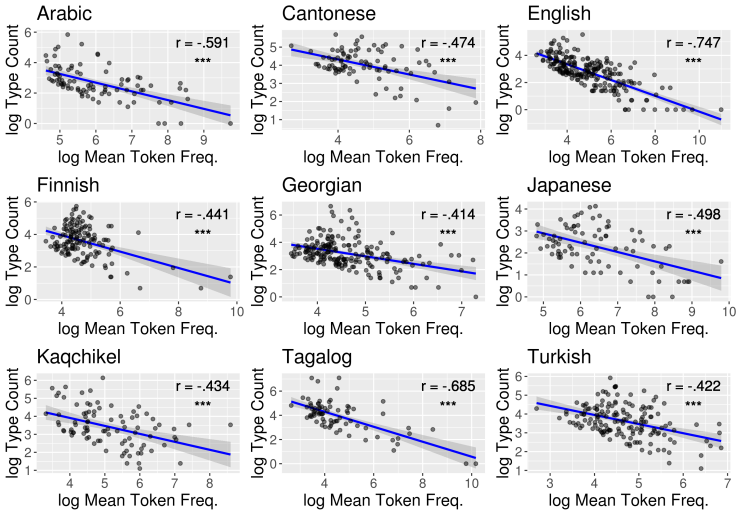


Figure 1. Inverse relation between type count and average frequency for word-initial biphone (2 segment sequences) in a subset of the 25 tested languages.

[1]This avoids effects of the so-called *Zipfian tail*, which creates an uninteresting positive relationship between type and token frequencies.

## References


Bentz, C., & Cancho, R. Ferrer-i. (2016). Zipf's law of abbreviation as a language universal. *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Prado Martin, F. Moscoso del. (2013). The missing baselines in arguments for the optimal efficiency of languages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*.

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

Zipf, G. K. (1935). *The psycho-biology of language*. Mifflin Houghton Publishing.