INCREMENTAL WORD PROCESSING HELPS SHAPE THE LEXICON

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

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

One of the most widely attested patterns in human language is that more frequent words tend to be shorter and less frequent words tend to be longer (Zipf, 1935). Further work showed that average probability of words in context (Piantadosi, Tily, & Gibson, 2011) is a better predictor of word length, updating Zipf's original observation. This pattern has been proposed to arise as result of pressures for languages to evolve to become more efficient communication systems: listeners need more information to accurately identify less probable words, and on average, a greater number of sounds in a word provides more information to a listener for word recognition. However, not all sounds in the word are uniformly informative for word recognition.

Listeners process spoken words incrementally, continually updating hypotheses about the identity of the word as they perceive each sound in sequence (McClelland & Elman, 1986; Norris & McQueen, 2008). As a consequence, earlier sounds in the word contribute more information on average to word recognition than later sounds, because they can, on average, rule out more possible alternatives (see Fig. 1). Here, we show that word that are less contextually probable are more likely to begin with highly informative sequences of sounds. Specifically, less predictable words are more likely to begin with sounds that rapidly distinguish the word from others in the lexicon. This is consistent with previous evidence that the lexicon is under pressure to evolve to serve as an efficient code, and further that as well as affecting the length of words, this pressure can affect the distribution of the sounds that make up the word with respect to the rest of the lexicon.

Methods and Results

We collected word frequencies in 5 languages: English, Dutch and German (Baayen, Piepenbrock, & Gulikers, 1995) and Japanese and Arabic (`CallHome` Corpus). We restricted our analysis to unaffixed words to avoid any potential confound with the effect of suffixes or prefixes on word processing. Following Son and Pols (2003) and Cohen-Priva (2015), we calculated the information of each phonetic segment of these words, as the -log probability of the segment, given the previous segments in the word, e.g. for the [f] in sphinx, the -log frequency of

DOI:10.12775/3991-1.050

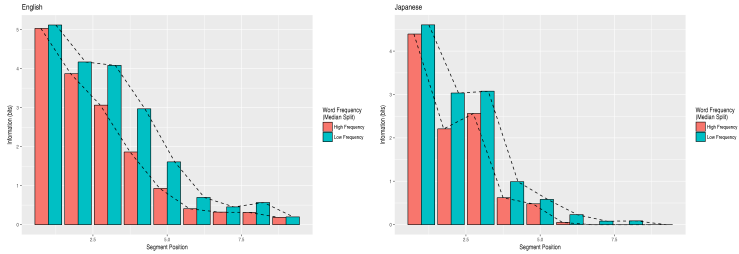

Figure 1. The average information provided by each segment position for English and Japanese words. Less frequent words contain more informative segments and this effect is much stronger early.

words with initial [sf] divided by words with the initial segment [s]. We calculated this in two ways: one including a measure of word frequency and one without.

We constructed a mixed-effect linear model to predict segmental information given position in the word and the word's frequency, with a random intercept for each word. Less frequent words contain significantly more informative sounds, independent of length for all languages. There was a significant interaction between position in word and word frequency, indicating that the effect of word frequency on segmental information is primarily found at the beginning of the word. We then fit a linear regression line to segmental information for each word. We found that less frequent words begin with significantly more informative segments. As a further control, we then compared the per-word regressions against those of a novel lexicon in which the order of all segments was reversed (see Fig. 2). We found that there was a much stronger effect of frequency in the unmodified lexicon.

Overall, our findings show that less expected contain more information in their sounds and this extra information is preferentially early. These results suggest that the lexicon has evolved to a state where the lexicon is partially optimized for listeners incremental processing of words.

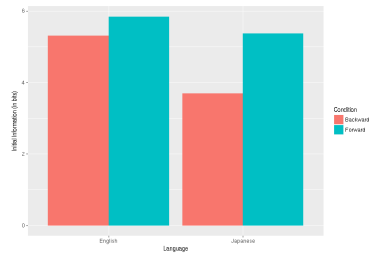

Figure 2. The initial information for words in English and Japanese, indicating this preference for early information is a product of the linear order of segments.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The celex lexical database (release 2): Linguistic data consortium. *University of Pennsylvania, Philadelphia, PA, USA*.

Cohen-Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Dick Goldhahn, T. E. . U. Q. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th international language ressources and evaluation (lrec'12).*

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive psychology*, *18*(1), 1–86.

Norris, D., & McQueen, J. M. (2008). Shortlist b: a bayesian model of continuous speech recognition. *Psychological review*, *115*(2), 357.

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.

Son, R. van, & Pols, L. C. (2003). How efficient is speech. In *Proceedings of the institute of phonetic sciences* (Vol. 25, pp. 171–184).

Zipf, G. K. (1935). The psycho-biology of language.