LANGUAGE-SPECIFIC CONSTRAINTS ON WORD FORM PREDICT SEGMENT INFORMATION ACROSS THE WORD

Andrew Wedel[1], Adam King[1], Adam Ussishkin[1]

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

King & Wedel (in press) found that each of 20 languages within a typologically-diverse set exhibited a significant inverse correlation between word probability and segmental information, that is, words which are less probable tend to contain segments that disambiguate from more lexical alternatives. They interpreted this in terms of Zipf's Law of Abbreviation (1949) which argued that the negative correlation between word probability and length arises from effort reduction on the one hand and maintainance of sufficient information in the signal on the other. In this case, King & Wedel argued that over time, greater phonetic reduction in higher probability words not only tends to promote loss of segments (Zipf 1949; Kanwal et al. 2017; Mahowald et al. 2018), but also leads to creation of more common segments and segment sequences, leading to relatively lower disambiguating information in the segments of high probablity words.

Lexical access in listeners proceeds incrementally as the speech stream is perceived, with the result that early segments in a word tend to provide more information than later segments (van Son & Pols 2003; Magnuson et al. 2007). On this basis, King & Wedel predicted that the correlation between word probability and segment information should evolve to be strongest at word beginnings, and decay late in the word. Using a measure of the difference in correlation between word probability and segment information early versus late in the word (see King & Wedel in press for method details), they found in fact that the majority of languages in the dataset did show the predicted pattern. However, a minority of languages, such as Hebrew and Arabic, did *not* show a significantly greater correlation early in the word. The question we address here is why some languages do not show this predicted pattern.

We test the hypothesis that the failure to show preferential optimization of segment information early in the word arises from language specific constraints on word formation which create a denser lexicon, that is, a lexicon in which words tend to be disambiguated from each other by fewer segments. As an example, the lexical meaning of words in Semitic languages like Hebrew and Arabic is largely carried by tri-consonantal roots. This restriction of word contrast to just three consonants means that for most words, all three consonants are required to disambiguate from alternatives (Ussishkin 2005). In contrast, languages like Georgian and Dutch have relatively large phoneme inventories, complex syllable structures, and allow variable word lengths, with the result that these lexicons tend to be sparser so later segments in the word tend to provide less information. Building on Wedel, Ussishkin & King (2019), we hypothesize that languages with sparser lexicons like Georgian and Dutch tolerate relatively greater reduction late in the word over time because those later segments are less likely to be informative to begin with, resulting in a pattern in which low probability words show their highest information segments at their beginnings. Conversely, when language specific constraints result in a more densely packed lexicon, segments across the word contribute more evenly to lexical disambiguation and so lower probability words show more evenly high segment information across the word.

In this study we use overall mean edit distance between words as a proxy measure for average lexicon density. As described above, King & Wedel showed that for most languages in the dataset, there was a significant interaction between word probability and a factor measuring bias of higher segment information toward the word-beginning, indicating that lower probability words tend to have a greater bias toward high early segment information. Here, we show that as hypothesized, mean edit distance itself significantly predicts the strength of this relationship, where low mean edit distance (e.g., as in Hebrew) is correlated with a lower bias toward early higher information (see Fig 1.) We have tested this relationship with a variety of other approaches and it remains robust. These findings contribute to the growing body of work in linguistic and cultural evolution on the influence of external constraints on the development of system-internal patterns.

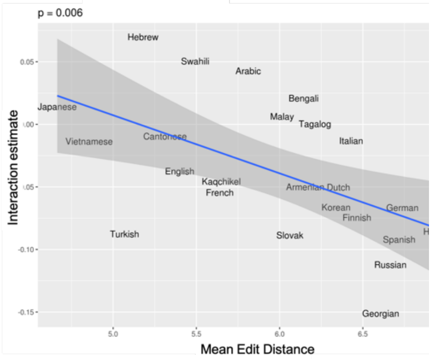

Figure 1. Correlation between mean edit distance and individual language model estimates for the interaction of word probability and early-bias for higher segment information. Languages with a higher mean edit distance (i.e, a sparser lexicon) are significantly more likely to show a bias toward early high segment information in lower probability words.

References

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165, 45–52.

King, A. & Wedel, A. (in press). Greater early disambiguating information for less probable words: the lexicon is shaped by incremental processing. *Open Minds.*

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(1), 133–156.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, 42(8), 3116–3134.

Ussishkin, A. (2005). A fixed prosodic theory of nonconcatenative templaticmorphology. *Natural Language & Linguistic Theory*, 23(1), 169–218.

van Son, R., & Pols, L. C. (2003). How efficient is speech. In *Proceedings of the Institute of Phonetic Sciences*, 25, 171–184.

Wedel, A., Ussishkin, A., & King, A. (2019). Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguistica*, 40(1), 231-248.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Welsey: Reading, Mass.