The Effect of Word Informativity on Word Shape

Psycholinguistic research have shown that listeners are able to incrementally process words as they are heard, progressively updating inferences about what word is intended as the phonetic signal unfolds in time (Allopenna et al. 1998, McMurray et al. 2002). For the majority of words in the lexicon, there is a point at which the word becomes uniquely identifiable before the final segment, the *uniqueness point*. Assuming accurate perception, segmental material after the uniqueness point is redundant. However, perceptual errors and contextual predictability create uncertainty. As a result, redundancy should contribute more to accurate lexical access when a word is less predictable.

Here, we show that words with greater average informativity possess relatively earlier uniqueness points. That is, more of the total length of the word is devoted to redundant material occurring after the uniqueness point. We operationalized word informativity as average forward trigram informativity (Piantadosi et al. 2011) in COCA (Davies 2009), and calculated uniqueness points using phonemic lemma representations with a frequency of greater than 1 per million based on the Carnegie Mellon Pronouncing Dictionary. Linear regression predicting the number of post-uniqueness point segments by average informativity and total word-length showed a significant negative correlation between average informativity and number of post-uniqueness point segments ($p = .001$).

Secondly, because words that are less probable in context begin their processing with effectively more competition, lexical access should benefit more from segmental cues in order to reduce the set of alternatives earlier in processing. As a result, we expect that words which are on average less contextually probable should evolve to contain more informative early segments and therefore earlier uniqueness points. We found that linear regression predicting the number of segments up to and including the uniqueness point by average informativity and total word-length showed a significant positive correlation ($p = .01$). These results are consistent with the notion that the distribution of word-internal segmental information, with respect to the set of existing alternatives, is modulated over the course of language evolution to support rapid and accurate lexical access.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4):419–439.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2):B33–B42.

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