Evaluating Lexical Frequency Measures for Sociolinguistic Variation

Ruaridh Purse and Meredith Tamminga
University of Pennsylvania

Word frequency has been demonstrated to be a robust predictor of sociolinguistic variation, such that frequent words behave differently from infrequent words (Pierrehumbert, 2002; Bybee, 2002). In response, sociolinguists will typically control for a measure of lexical frequency in their analyses of such phenomena. However, the concept of lexical frequency can be captured with a number of different measures, and the appropriate measure to use remains an open question (cf. Hay, 2001; Walker, 2012). The present study investigates the capacity for 3 different measures of frequency to account for variance for two morphophonological variables: TD (e.g. *old* vs. *ol'*) and ING (*working* vs. *workin'*).

11964 tokens of words with an underlying word-final coronal stop (TD), and 5452 tokens of verbs in the present participial or gerundive forms, i.e. with suffixal *-ing* (ING), were collected from a sample of 118 sociolinguistic interviews with white speakers from the Philadelphia Neighborhood Corpus (PNC) (Labov & Rosenfelder, 2011). Each token was manually coded for the variants of TD and ING, as well as morphological and phonological context and speaker demographic information.

Tokens were also coded for three measures of lexical frequency according to the $\text{SUBTLEX}_{\text{US}}$ database (Brysbaert & New, 2009). For Wholeword frequency – typically used in sociolinguistics – each unique orthographic string is assigned a separate frequency value regardless of phonological, morphological, or semantic relations. This means that each of *dog*, *dogs*, *read*, and *reed* receive a different value, while *left* (direction) and *left* (past of *leave*) receive the same value. For Root frequency, the frequency of each item is calculated as the sum of Wholeword frequencies sharing its stem. Finally, Conditional frequency is the proportion of an item's parent Root that is constituted by a particular Wholeword, and represents the frequency of a Wholeword given the Root. These measures are not highly correlated with each other. Mixed-effects logistic regression models controlling for grammatical class and speech rate contained each possible combination of the three measures. Then, nested models were compared for degree of optimisation with the addition of each measure to each model. Relevant statistics for this were Aikaike and Bayesian information criteria, and likelihood ratio tests. For TD, all frequency measures predict TD outcomes as expected, but the data is best captured by Root frequency, in contrast with the findings of Brysbaert and New (2009). This measure significantly improved each model, regardless of what other measures were present. However closer inspection revealed that this effect was only present in monomorphemes and not words with *-ed* suffixes (Figure 1). The case of ING is more complicated. Conditional frequency is correlated with the variable in an unexpected direction, and all measures significantly improve most models by their inclusion. The most significant and consistent in this respect is Wholeword frequency, the effect of which is observed in both gerundive and progressive forms (Figure 2).

Our findings suggest an opportunity for further exploration of our conception of lexical frequency and the role it plays in sociolinguistics. For some phenomena, Wholeword frequency may play a key role, but even in such cases it does not appear to account for all of the data.

References

Brysbaert, M. & New, B. 2009. Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure of American English. *Behavior Research Methods, 41*(4), 977–990.

Bybee, J. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change 14*, 261–290.

Hay, J. 2001. Lexical frequency in morphology: is everything relative? *Linguistics, 39*(6), 1041– 1070.

Labov, W. & Rosenfelder, I. 2011. *The Philadelphia Neighborhood Corpus*. Philadelphia: University of Pennsylvania. Online: http://fave.ling.upenn.edu/pnc.html

Pierrehumbert, J. B. 2002. Word-specific phonetics. In: *Laboratory Phonology VII*. Berlin: Mouton de Gruyter 101–139.

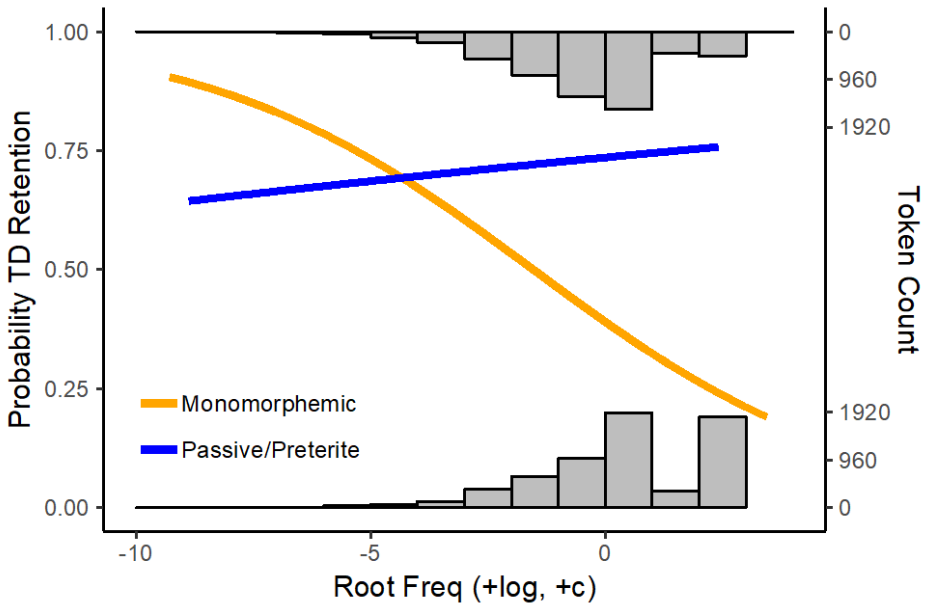

Figure 1. Histograms for TD by Root frequency, log-transformed and centred. Simple logistic Regression curves for each grammatical class.

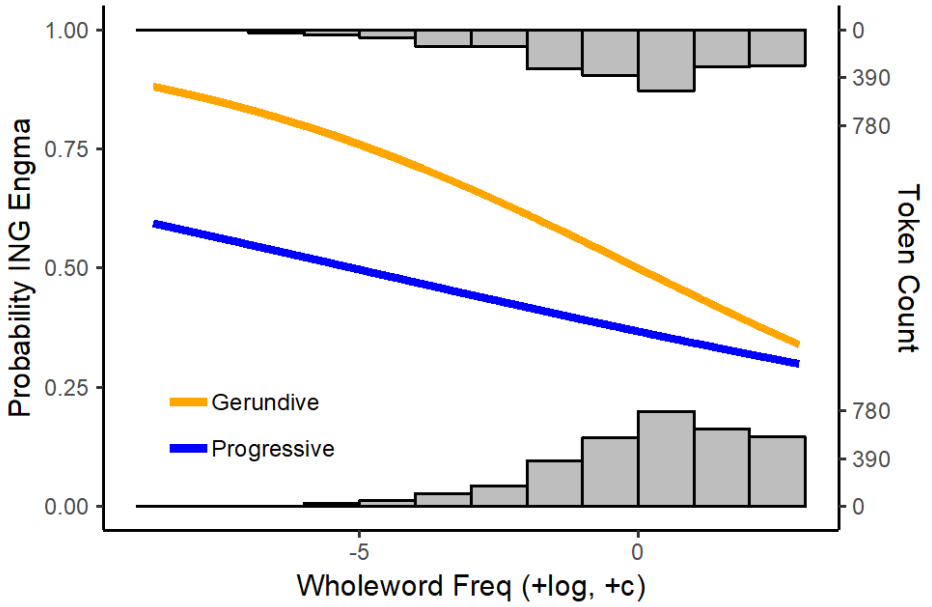

Figure 2. Histograms for ING by Wholeword frequency, log-transformed and centred. Simple logistic regression curves for each grammatical class.