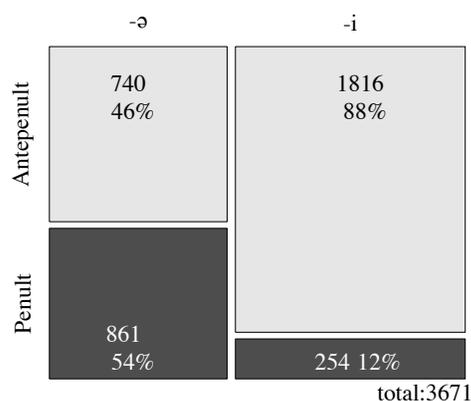


The phonological grammar is probabilistic: New evidence pitting abstract representation against analogy

Claire Moore-Cantwell, Yale University

Speakers and listeners extend both categorical and probabilistic regularities in the lexicon of their native language to novel forms. Ernestus and Baayen (2003); Hayes et al. (2009) demonstrate that speakers can ‘probability match’ - rather than applying a trend in the lexicon categorically to new forms, speakers produce a distribution of output forms which matches the distribution of form types found in the lexicon. As Ernestus and Baayen demonstrate, this probability matching behavior can be modeled equally well via a set of abstract generalizations situated within a probabilistic grammar, or via a process such as analogy which is an epiphenomenon of the organization of the lexicon. Experimental evidence such as Guion et al. (2003) suggests that both mechanisms are at work, a notion that is formalized in dual-route or two-systems models (Ullman, 2004; Pinker, 1999). These models typically incorporate abstract grammatical knowledge for categorical phenomena, and analogical mechanisms for probabilistic phenomena.

I examine a probabilistic trend within the English stress system, showing that speakers extend it to new words, but they do not use information about particular existing words to do so. I argue that speakers’ knowledge of this trend is both abstract and probabilistic in nature. This supports the use of inherently probabilistic grammatical models such as Maximum Entropy to model probability matching behavior (Goldwater and Johnson, 2003; Hayes and Wilson, 2008; Coetzee and Kawahara, 2013).

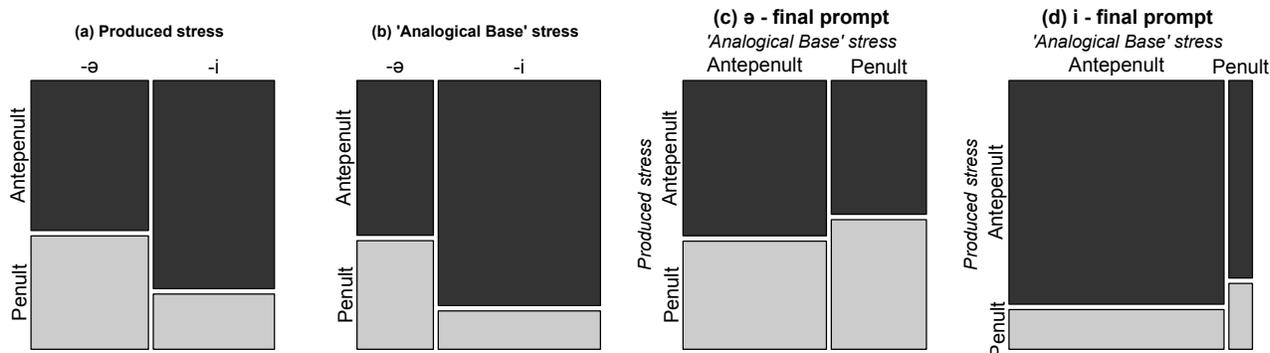


The probabilistic trend: In English words longer than two syllables, stress is typically penultimate (‘banána’) or antepenultimate (‘Cánada’). A search of the CMU pronouncing dictionary (Weide, 1994) revealed that [i]-final words were biased towards taking antepenultimate stress, and [ə]-final words were unbiased. In words at least 3 syllables long, 88% of i-final words were antepenultimately stressed, but only 54% of ə-final words, were antepenultimately stressed. This trend can be captured in a constraint-based phonological grammar through the use of a constraint which demands that a final [i] be extrametrical.

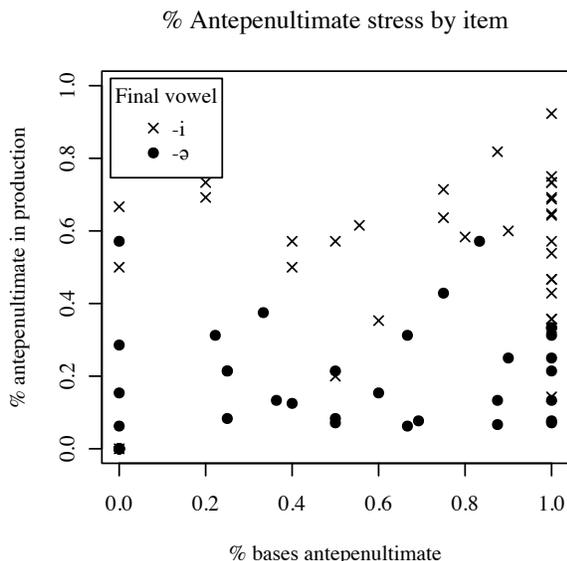
Methods: Building on the methods of Guion et al. (2003), 50 participants recruited through Amazon Mechanical Turk performed 2 tasks. **PRODUCTION TASK:** Nonwords (half -i, half -ə) were constructed so as to have very sparse neighborhoods (less than 0.01) according to the Generalized Neighborhood Model (Bailey and Hahn, 2001). Nonwords were presented auditorily as three individual syllables, each spoken as a separate prosodic word ([bæ] [mæ] [ki]). The syllables were resynthesized so that they had identical acoustic cues to stress: duration, intensity, and pitch contour. Participants were recorded as they spoke the syllables fluently as a single word. Next, participants ‘transcribed’ their own production by listening to 2 versions of the nonword ([bæmæki], [bəmæki]) and selected the version most similar to what they produced. **ANALOGICAL BASE TASK:** Participants heard each stress-ambiguous

nonword again, and filled in a blank with a real word that it reminded them of.

Results: Data from 32 participants was analyzed, all at least 90% accurate in their ‘transcriptions’. Participants extended the probabilistic trend in the lexicon to nonwords: i-final words took antepenultimate stress 77% of the time (88% in the lexicon) while ə-final words took antepenultimate stress only 58% of the time (54% in the lexicon). In production, i-final nonwords received more antepenultimate stress than ə-final nonwords (a). Likewise, i-final ‘analogical bases’ provided by participants were more likely to be antepenultimately stressed than ə-final bases (b). However, the stress of these analogical bases did not directly relate to a participant’s produced stress (c,d).



A mixed effects logistic regression (random slopes and intercepts for subjects and items) showed an effect of final vowel on produced stress (-ə items have less antepenultimate stress than -i, $\beta = -1.27$, $p < 0.001$, $AIC = 290$). The stress of the analogical base provided by each participant did not predict that participant’s produced stress, and did not improve the model’s fit (Penult vs. Antepenult, $\beta = 0.42$, $p = 0.20$, $AIC = 290$).



Analogical base responses to each nonword were also examined in aggregate. Nonwords differed from each other in the distribution of analogical bases given. Some were majority antepenultimately stressed, and others were majority penultimately stressed. Each item’s percentage of antepenultimately stressed bases was calculated, and is plotted here against the item’s rate of antepenultimate stress in production. The two percentages are not related ($\rho_i = 0.16, \rho_ə = 0.37$). Participants successfully extended the trend in the lexicon to nonwords, and their chosen analogical bases follow the trend in aggregate, but these two behaviors do not proceed from the same underlying (lexical access) process.

Participants ‘probability-match’ the trend in the lexicon for i-final words to take antepenultimate stress, but this behavior is not attributable to an analogy process. Rather, speakers’ phonological grammar must be able to represent probabilistic tendencies as well as categorical generalizations.