# Similarity, feature-based generalization, and bias in novel onset clusters

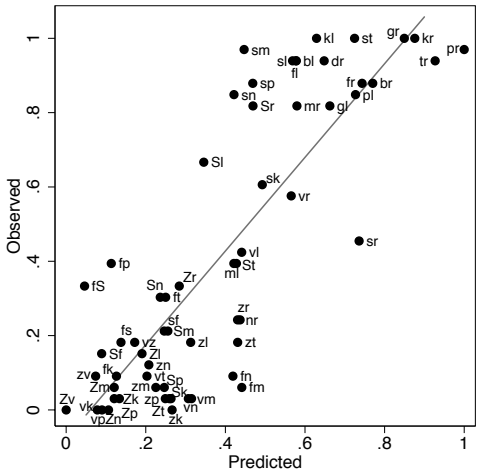
Adam Albright, MIT

The past decade has seen significant advances in modeling gradient preferences of English speakers using grammars with probabilistic rules, constraint rankings, or weights inferred from the statistical properties of the lexicon (Coleman and Pierrehumbert 1997; Treiman, Kessler, Knewasser, Tincoff, and Bowman 2000; Frisch, Large, and Pisoni 2000; Bailey and Hahn 2001; Hammond 2004; Hayes and Wilson, in press; and many others). In study after study, where lexical counts have shown probabilistic differences between different cluster types, experiments with native speakers have shown corresponding behavior preferences. At the same, studies investigating substantive biases in phonology have begun to focus on cases in which speakers' preferences do not appear to follow straightforwardly from lexical statistics (Moreton 2002, 2007; Wilson 2003, 2006; Zhang and Lai 2006; Berent et al, in press). For example, Berent et al. (in press) show that English speakers prefer initial *#bn* over *#bd*, even though few or no English words begin with either cluster. These results show that speaker preferences cannot be guided solely by obvious lexical statistics such as literal biphone frequency. However, they leave open the question of whether such preferences might be learned by some less direct means, perhaps involving similarity, phonological features, or some other mechanism for generalizing from attested to unattested sequeces. In this talk, I report on some attempts to take on this question, exploring the capacity of similarity-based and natural class-based models to predict generalization to novel clusters.

Broadly speaking, there are two fundamentally different ways that knowledge of existing sequences could be leveraged to generalize to novel words: by gauging activation of similar exemplars in the lexicon, or by abstracting a grammar that encodes knowledge about possible sequences (Bailey and Hahn 2001). In order to test the predictions of a similarity-based exemplar model, I use an implementation of Bailey and Hahn's Generalized Neighborhood Model, which assesses the aggregate similarity of novel items to the lexicon of existing words. As a test of feature-based generalization, I consider the predictions of a model that keeps track of n-grams of natural classes (Albright, in prep). Before we can test the predictions of the models for preferences among unattested sequences, it is first useful to check their predictions for legal sequences, since work on gradient phonotactics suggests that lexical models should do well in modeling preferences for common vs. rare attested combinations. When the models are tested on a set of 70 randomly selected wug words and their predictions are compared to native speaker ratings, we find moderately strong correlations ($r$ between .45 and .60, depending on the precise parameters and training set employed). This confirms that both the similarity based and natural class based models have a certain degree of plausibility as models of how speakers generalize gradient preferences among attested sequences. Furthermore, Hayes and Wilson (in press) provide evidence suggesting that this result may extend well to unattested sequences: they compare the predictions of their inductive model against data involving acceptability ratings from seventh graders (Scholes 1966) for a mix of attested and unattested sequences, showing that the model achieves a strikingly close fit to human ratings. When the current models are trained and tested in a similar fashion, they likewise show a reasonably good fit; the performance of the natural class model is shown in (1) ($r(60) = .83$). Thus, it seems promising that such models could learn preferences such as *#bn* $\succ$ *#bd*.

The predictions of the GNM and natural class based models, as well as Hayes and Wilson's MaxEnt model, were compared for a variety of unattested onset clusters for which judgment data is available. The results show that for some comparisons, all three models agree that there is greater lexical support for the preferred cluster. For example, *#bn* receives greater support from existing onsets than *#bd* does, no matter whether it is evaluated via combinations of natural classes or by raw perceptual similarity ((2)). Although this result rests on a particular set of assumptions about how the models are trained and tested, it does raise the possibility that the preference documented by Berent et al. could somehow be inferred from the data of English. On the other hand, all three models show a preference for *#bn* $\succ$ *#bw*, while speakers show exactly the opposite preference (Albright 2007). Similarly, significant preferences for *#bw* $\succ$ *#dl* (Moreton 2002) and *#sp* $\succ$ *#sk* (Scholes 1966) are not correctly predicted by any of the models tested ((3)–(5)). Although it is not clear to what extent these discrepancies point to failings of the model and to what extent they diagnose substantive biases, it is perhaps not a coincidental that they all mirror differences in cue availability for $C_1$ or $C_2$ in the cluster. I show that when phonetically motivated biases are imposed on the inductive model, it can achieve a very close fit to human ratings. I conclude that the search for evidence for substantive biases must be a multi-pronged effort involving not only experimentation, but also a good faith attempt to model preferences based on linguistic experience and a careful consideration of the nature of the biases involved.

(1) Predictions of natural class model for Scholes (1966) data




(2) *bn* $\succ$ *bd*: Potentially learnable

- $P(\begin{bmatrix} +\text{labial} \\ -\text{sonorant} \\ -\text{continuant} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{sonorant} \end{bmatrix}) > P(\begin{bmatrix} -\text{sonorant} \\ +\text{labial} \\ -\text{continuant} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{voice} \end{bmatrix})$
- Exemplar model: existing *bl* more similar to *bn* than *bd*
- Hayes and Wilson MaxEnt model: *#bn* $\succ$ *#bd* (but **not** *#pn* $\succ$ *#pt*)

(3) *bw* $\succ$ *bn* (Albright 2007): Not learned

- $P(\begin{bmatrix} +\text{labial} \\ -\text{sonorant} \\ -\text{continuant} \end{bmatrix}\begin{bmatrix} +\text{labial} \\ +\text{sonorant} \end{bmatrix}) < P(\begin{bmatrix} +\text{labial} \\ -\text{sonorant} \\ -\text{continuant} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{sonorant} \end{bmatrix})$

(4) *bw* $\succ$ *dl* (Moreton 2002): Not learned

- $P(\begin{bmatrix} +\text{labial} \\ -\text{sonorant} \\ -\text{continuant} \end{bmatrix}\begin{bmatrix} +\text{labial} \\ +\text{sonorant} \end{bmatrix}) < P(\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \end{bmatrix}[+\text{approx}])$

(5) *sp* $\succ$ *sk* (Scholes 1966): Not learned

- $P(\begin{bmatrix} +\text{coronal} \\ +\text{continuant} \\ -\text{voice} \end{bmatrix}\begin{bmatrix} +\text{labial} \\ -\text{continuant} \\ -\text{voice} \end{bmatrix}) \approx P(\begin{bmatrix} +\text{coronal} \\ +\text{continuant} \\ -\text{voice} \end{bmatrix}\begin{bmatrix} +\text{dorsal} \\ -\text{voice} \end{bmatrix})$

## References


Bailey, T. and U. Hahn (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language 44*, 568–591.

Berent, I., D. Steriade, T. Lennertz, and V. Vaknin (to appear). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*.

Coleman, J. S. and J. Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*, pp. 49–56. Somerset, NJ: Association for Computational Linguistics.

Frisch, S. A., N. R. Large, and D. B. Pisoni (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language 42*, 481–496.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan, and G. Miller (Eds.), *Linguistic Theory and Psychological Reality.*, pp. 294–303. MIT Press.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies 4*, 1–24.

Hayes, B. and C. Wilson (to appear). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition 84*, 55–71.

Moreton, E. (2007). Learning bias as a factor in phonological typology. Paper presented at the 26th meeting of the West Coast Conference on Formal Linguistics, Berkeley, California, April 27-29.

Scholes, R. J. (1966). *Phonotactic Grammaticality*. Janua Linguarum. The Hague: Mouton.

Treiman, R., B. Kessler, S. Knewasser, R. Tincoff, and M. Bowman (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe and J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pp. 269–282. Cambridge: Cambridge University Press.

Wilson, C. (2003). Experimental investigation of phonological naturalness. In G. Garding and M. Tsujimura (Eds.), *WCCFL 22 Proceedings*, Somerville, MA, pp. 533–546. Cascadilla Press.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science: A Multidisciplinary Journal 30*, 945–982.

Zhang, J. and Y. Lai (2006). Testing the role of phonetic naturalness in Mandarin tone sandhi. *Kansas Working Papers in Linguistics 28*, 65–126.