Natural classes are not enough: Biased generalization in novel onset clusters

Adam Albright
MIT

1 Introduction

It has long been recognized that a powerful source of information about the phonological knowledge that speakers have comes from the fact that they can generalize this knowledge to novel items. A classic example of this, discussed by Chomsky and Halle (1965) and recited to countless undergraduate classes and strangers on airplanes ever since, is the fact that native speakers typically judge *blick* [blɪk] to be a possible word of English, but tend to agree that *bnick* [bnɪk] would not be possible. Although **blick* is perhaps the most famous example of generalization to a novel string, it is unfortunately not all that revealing about the form that the relevant knowledge takes or how speakers acquire it, since all that would be required in order to reject **bnick* and favor *blick* is a quick look at the lexicon: there are no English words starting with *#bn*. A slightly more telling fact is that speakers show preferences for some attested sequences over others (*blick* [blɪk] ≻ ?*thwick* [θwɪk]). This particular preference could follow straightforwardly from the fact that there are relatively few English words starting with [θw] (*thwart*, *thwack*), but at least it shows that speakers have more refined knowledge than simply whether there are any existing words that start the same way.

A much more interesting kind of fact is when speakers prefer one unattested sequence over another: **bnick* [bnɪk] ≻ ***bdick* [bdɪk], ***bzick* [bzɪk]. In such cases, the preference that we observe could not be due to the fact that there are more *#bn* words than *#bz* or *#bd* words, since there are no words that begin with any of these clusters. Ultimately, what we would like to know is to what extent speakers' preferences are learned (directly or indirectly) from the data of English, and to what extent they reflect prior, universal biases.[1] There is a growing body of literature that attempts to document substantive biases by observing unmotivated preferences for one pattern over another equally attested pattern (Zhang and Lai 2006; Becker, Ketrez and Nevins 2007; Moreton, to appear), or by isolating preferences for some unseen items over others (Wilson 2003, 2006; Moreton 2007; Berent, Steriade, Lennertz, and Vaknin, to appear). Such works challenge the idea that all phonological preferences are based on direct experience with obvious statistical

[1] I use the terms *universal* and *prior* to mean "independent of linguistic data". The term is neutral with respect to the question of what aspects of these biases are innate, and what aspects are inferred via universally available phonetic experience (i.e., by *inductive grounding*; Hayes 1999, Hayes and Steriade 2004).

properties of one's native language. However, they generally leave open the possibility that the observed preferences may have been inferred in some less obvious indirect fashion. Concretely, it seems possible that any preference English speakers show for *#bn* over *#bd* could be based on native language experience with stop + sonorant sequences (such as *#br*, *#bl*), which, though not involving *#bn* directly, provide indirect evidence that structures like *#bn* might be conceivable.

The current paper explores this issue, asking to what extent preferences for certain novel onset clusters could in principle be learned from the input data. Of course, even if a preference for *#bn* $\succ$ *#bd* can be inferred from statistical properties of English, there is no guarantee that speakers actually do require data to learn it; however, the interest of the case would be diminished as a straightforward example of substantive bias. The more interesting outcome is if the preference is evidently not learnable using general purpose learning strategies. In this case, we are in a position to construct a argument from the poverty of the stimulus in favor of universal phonotactic knowledge. In accordance with Pullum and Scholz (2002), such an argument would have the following form: first, we specify the ACQUIRENDUM, or the grammatical knowledge imputed to native speakers—here, a constraint ranking in which stop+nasal sequences are preferred over stop+obstruent sequences, as in (1).

(1) Acquirendum: grammatical ranking preferring *#bn* $\succ$ *#bd*

*stop/__ obstruent $\gg$ *stop/__ nasal

The next step in the argument is to provide evidence in support of the claim that speakers have acquired the acquirendum. For example, we could point to the fact that English speakers tend to perceive *#bdVC* sequences as disyllabic more often than they perceive *#bnVC* as disyllabic (Berent, Steriade, Lennertz, and Vaknin, to appear), indicating a greater unwillingness to treat *#bd* as an onset cluster. Alternatively, we might show that native speakers rate *#bn*-initial words as more acceptable or plausible as English words; data to this effect will be presented in section 2. Next, we enumerate the types of linguistic data that would allow learners to infer the acquirendum (if such data were available). This could be very direct evidence, such as greater numbers of existing words starting with *#bn* than with *#bd*, but it could involve less direct inference, perhaps from the rate of attestation of word-medial *bn* vs. *bd*, or an greater reluctance to avoid deriving *bd* through processes like deletion of unstressed schwa (*banana* /bənǽnə/ $\rightarrow$ [bnænə], vs. hypothetical *badana* /bədǽnə/ $\not\rightarrow$ *[bdænə]). We must then show that such comparisons are not actually available to children acquiring English (= INACCESSIBILITY). For example, as noted above, English has no words starting with either *#bn* or *#bd*[2], and medial *bn* and *bd* are both uncommon—in fact *bn* is slightly more so.[3] Furthermore, although words like *banana* may provide evidence for initial /#bən/ $\rightarrow$ [#bn], there are no initial /#bəd/ words to show the needed comparison of relative reluctance to derive /#bd/,[4] and what few medial /bən/ and /bəd/ words there are tend not to syncopate (*eb**o**ny*, *deb**o**nair*, *Leb**a**non*; *nob**o**dy*). Thus, there does not appear to be any clear evidence for *#bn*

[2]Unless you happen to be in a speech community that uses words like *bnai brith* or *bdellium*.

[3]Five most frequent in CELEX: [bn] *abnormal*, *obnoxious*, *subnormal*, *drabness*, *hobnailed*; [bd] *abdomen*, *subdued*, *abdication*, *abduction*, *subdivision*.

[4]Some possible points of contention: *bedevil*, *bedazzle*, *bidet*. Even if these words do have [ə], however, they are so infrequent that they presumably rarely undergo syncope.

$\succ$ *#bd* based on the occurrence of [#bn] or the avoidance of [#bd]. Finally, we must certify that the relevant data is not only unavailable, but it is also absolutely essential to learning (= INDISPENSABILITY). If these conditions are true and the acquirendum could not have been learned from linguistic data, then we must conclude that it reflects a universal preference.

Arguments of this form are notoriously difficult to construct, and one might question various aspects of it. The first point that deserves attention is the acquisition evidence. In point of fact, although generalization to novel strings has long been a part of the rhetoric of the study of phonology as a cognitive system, in practice there are remarkably few studies demonstrating systematic differences between unattested clusters (a point also noted by Berent et al.). In section 2, I will provide experimental evidence from acceptability ratings of non-words, showing that closely matched clusters do indeed fall along a scale of acceptability (*#bw* $\succ$ *#bn* $\succ$ *#bz*, *#bd*). Another area in need of scrutiny is the claim of indispensability: even in the absence of a grammatical preference for stops to occur in more sonorous contexts, couldn't a preference for *#bn* $\succ$ *#bd* arise via some other, non-grammatical comparison? In section 3, I will consider two similarity-based models, testing the extent to which they can predict the observed preferences solely on the basis of perceptual similarity to existing items. The first is an analogical model, which attempts to explain preferences for words like *bwick* or *bneed* based on their similarity to existing words like *brick* and *bleed*. The second is a model that focuses on perceptual similarity of novel clusters to existing ones—e.g., *#bw* and *#bn* sound better than *#bd*, *#bz* becase [w] and [n] are perceptually closer to *l*, *r*. As we will see, neither mechanism is sufficient to explain the observed preferences, bolstering the claim of indispensability for a grammatical preference.

The conclusion that non-grammatical similarity-based preferences are insufficient does not immediately motivate a need for prior biases, however, since it still leaves open the question of whether there might be a model of grammatical learning that could predict the observed preference based on the data of English. Indeed, Hayes and Wilson (to appear) provide an inductive model of constraint learning that performs extremely well in modeling data from a range of attested and unattested onset clusters. The data that they consider, taken from Scholes 1966, does not contain comparisons of clusters with stops + non-liquids (i.e., it contains *#br* and *#bl*, but not *#bn*, *#bd*, *#bz*), but it does contain comparisons among other unattested clusters such as *#sr* $\succ$ *#zr*. The high level of performance that their model achieves on these comparisons makes it seem promising that preferences like *#bn* $>$ *#bd* could also be learned. In section 4, I sketch a learning model that inductively posits constraints based on linguistic data, designed to test whether it is possible to support generalization to *#bn* based on natural classes (*#bl*, *#br* $\Rightarrow$ *#b*+sonorant). The claim of this section will be that although this model makes significant headway in predicting gradient differences in acceptability among attested sequences and also among some unattested clusters (such as those tested by Scholes), it turns out that good performance on these tasks does not guarantee that the model will distinguish correctly between novel clusters like *#bn* and *#bw*. Finally, in section 5, I show that the best available model for the data is one that incorporates both inductively learned constraints (reflecting statistical properties of English) and also prior constraints (reflecting a universal preference for stops to be followed by more sonorous segments).

2 The data

Many previous studies have investigated the acceptability of clusters using novel words (Greenberg and Jenkins 1964; Scholes 1966; Pertz and Bever 1975; Coleman and Pierrehumbert 1997; Hay, Pierrehumbert, and Beckman 2004; Moreton 2002; Davidson 2006; Berent, Steriade, Lennertz, and Vaknin, in press; Haunz 2007). The goals of the present study were (1) to gather ratings of closely matched unattested clusters alongside a wide range of existing or well-formed structures, and (2) to collect simultaneous repetition and ratings data for non-words. Data about a wide range of structures is important in allowing us to calibrate models of gradient preferences, and by pairing each novel onset cluster with multiple rhymes, we facilitate inferences beyond the particular set of nonce words in the experiment. Simultaneous repetition and ratings data provide a check that subjects were rating the intended items, and not misheard variants (an especially important consideration in the case of illegal or novel clusters). Furthermore, error analysis of repetitions can provide valuable additional sources of data such as the error rate across different clusters, the nature of repairs, and so on.

2.1 Experimental details

2.1.1 Stimuli

A set of 30 monosyllabic nonwords starting with *p-*, *b*-initial clusters was constructed by pairing onset clusters with a selection of different rhymes. Onset clusters included *#pl*, *#bl*, *#pw*, *#bw*, *#pn*, *#bn*, *#pt*, *#bd*, *#ps*, and *#bz*. Data from *#ps* will not be considered here, since subjects generally had a difficult time perceiving and repeating it accurately. Rhymes were chosen to control as much as possible for neighborhood density (as measured by aggregate lexical similarity counts; Bailey and Hahn 2001) and for bigram probability (as measured both by average bigram transitional probability in the word, and a natural class-based bigram model described in section 4). The result was a set of words like *bwadd* [bwæd] *bneen* [bniːn], and *bduss* [bdʌs]; the full set is in the appendix. (Novel words are given here with regular English orthography for expository purposes only—no written materials were presented to experimental subjects.) In addition, 170 filler items were included, containing a mix of items from previous nonce word studies (to facilitate comparison across studies) and words with other cluster types, not considered here.

Five of the fillers involved initial *#pr*, *#br*: *prundge* [prʌnd͡ʒ], *prupt* [prʌpt], *presp* [prɛsp], *brelth* [brɛlθ], *brenth* [brɛnθ], and five involved other clusters of interest (*blig* [blɪg], *blemp* [blɛmp], *pwist* [pwɪst], *pwuds* [pwʌdz], *ptep* [ptɛp]). Since these items also generally lacked close lexical neighbors, data from them will be included in the analysis; however, we must bear in mind that the *#pr*, *#br* words in particular involved much lower probability rhymes than the remaining clusters.

From among the 170 filler items, 70 items with no overt phonotactic violations (i.e., without unattested clusters) were chosen pseudo-randomly for purposes of calibrating the statistical models to be discussed below. These items were chosen from among the larger set, attempting to achieve as close to normal a distribution of subject ratings as possible. (For discussion of the importance

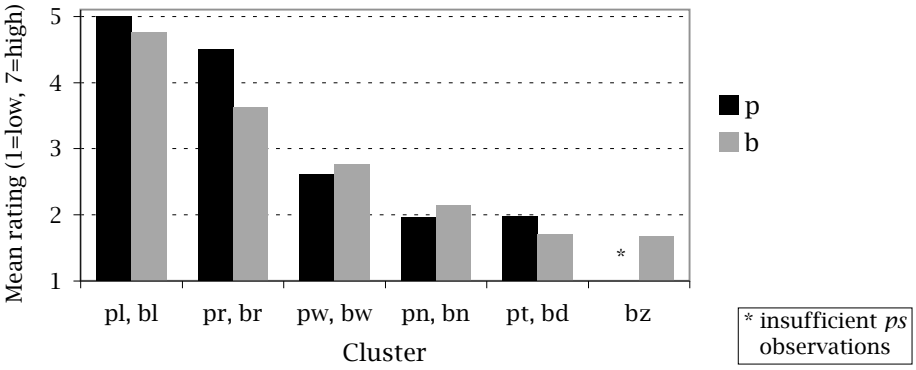


Figure 1: Preference for C_2: *bl* (>) *br* > *bw* > *bn* > *bd*, *bz*

of testing models against randomly sampled items, see Bailey and Hahn 2001.) Examples include *glack* [glæk], *pleen* [pliːn], *trusp* [trʌsp], *flet* [flɛt], *smat* [smæt], *stilf* [stɪlf], and *cobe* [koʊb].

The items were read by a phonetically trained male native speaker of English in simple carrier sentences, in which the nonword acted either as a noun or as a verb: "[bwæd]. *I like to* [bwæd]", or "[bwæd]. *This is a* [bwæd]". The recorded stimuli were checked to ensure that initial stops in C_1C_2 onset clusters had bursts cueing their presence, and no vocalic period between the burst and C_2 (defined as a voiced interval with clear formant structure distinct from the following consonant). Stimuli were RMS equalized in Praat, using a script by Gabriël Beckers.[5]

2.2 Procedure

Novel words were presented in their frame sentences using PsyScope (Cohen, MacWhinney, Flatt, and Provost 1993). Presentations as nouns and as verbs were counterbalanced across subjects, and presentation order was randomized on a subject-by-subject basis. Subjects were instructed to repeat the novel word aloud, and then use the keyboard to enter their rating on a scale from 1 ("Completely impossible as an English word") to 7 ("Would make a fine English word"). Spoken responses were transcribed by two phonetically trained listeners, and if the listeners did not agree that the subject had repeated a particular word as intended, the rating from that trial was discarded.

2.3 Results

A preliminary analysis revealed no significant effect of part of speech, so ratings from noun and verb presentations were combined for subsequent analyses. Figure 1 shows that among the items with {*p*,*b*}-initial onset clusters, a clear ordered preference for C_2 was observed: *bl* (>) *br* > *bw* > *bn* > *bd*, *bz*. As might be expected, there was a clear preference for attested clusters (*bl*, *bl*, *pr*, *br*) over novel clusters. (Recall that the relative preference for *Cl* > *Cr* is most likely an experimental

[5] http://www.gbeckers.nl/pages/praatscripts.html

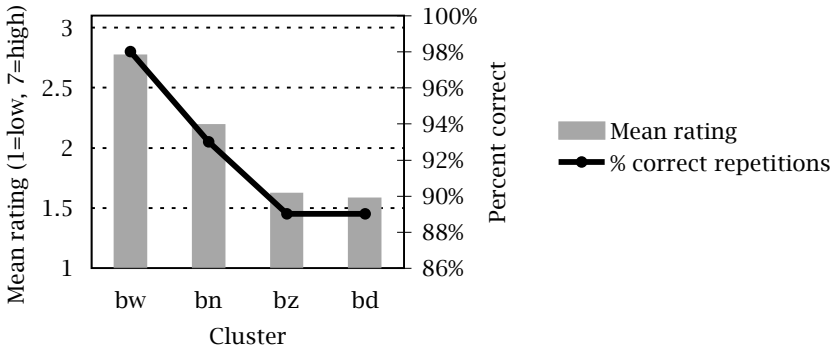


Figure 2: Relation between acceptability ratings and repetition accuracy

artifact due to the fact that these words contained less probable rhymes.) Among the novel onset clusters, a general preference can be observed for stops to occur before more sonorous elements ($w \succ n \succ d, z$). This is seen most clearly in the preference for *#pw*, *#bw* $\succ$ *#pn*, *#bn*, but is reflected also in the preference for *#bn* $\succ$ *#bd*, *#bz*. The preference for *#bn* $\succ$ *#bd* is seen also in the results of Berent et al. (in press), who found that English-speaking subjects more often judged *#bd*-initial words to be disyllabic. The preference is clearer for *b*C clusters than for *p*C clusters, which involved less data and did not show a significant differentiation of *pn* and *pt*.

Lower ratings were also mirrored by greater numbers of incorrect repetitions, shown for *b*C clusters in Figure 2. Taken together, this data confirms the preference for stops to occur before more sonorous elements.

This result establishes more detailed evidence that English speakers do in fact preferentially generalize to stop+consonant clusters with more sonorous second consonants. It also provides a set of quantitative ratings for individual novel words that can be used to test how well various models simulate these preferences with or without an explicit prior bias for certain types of clusters. Of course, an adequate explanation of preferences among novel clusters based on general purpose statistical learning principles must work not just for items with novel clusters, but must also do well in general on arbitrarily chosen novel sequences. In the next two sections, I will present possible statistical accounts of gradient generalization to novel items, benchmark their performance against the set of 70 "calibration" items, and then test their ability to model novel cluster preferences.

3 Similarity-based models

The first line of explanation to be considered here is that the gradient preferences documented in the previous section have nothing at all to do with the grammar of English, but rather involve a non-grammatical evaluation of the perceptual similarity of words with novel clusters to existing English words. This perceptual comparison could be an automatic one carried out in the course of mapping the acoustic signal onto linguistic categories, or it could be a task-dependent strategy. In either case, the possibility of a similarity-based account threatens the claim that a particular type

of grammatical knowledge is indispensable as a basis for differentiating unattested clusters.

3.1 An analogical account

It is a plausible and very widely held assumption that gradient acceptability judgments in phonology reflect two kinds of knowledge: on the one hand, there is grammatical knowledge, which assesses the legality of the novel sequence as a possible word. On the other hand, there is lexical knowledge, which allows one to assess the probability that the novel string would actually be a word. Even if it were the case that the grammar of English did not differentiate between *#bw*, *#bn* and *#bd* clusters (classifying them all as ungrammatical), it might nonetheless be possible to distinguish among them by virtue of their overlap with attested clusters, since there are many existing words with #C*w* onsets, fewer with #C*n* onsets, and none with #C*d* onsets. The hypothesis, then, is that words with certain clusters may receive greater analogical support from the lexicon.

Support from the lexicon for a novel word is typically calculated based on the degree of similarity between the new word and the set of existing words. Similarity between the novel word and a particular existing word is assessed by determining the number of transformations required to turn the nonce word into the real word (Greenberg and Jenkins 1964; Coltheart, Davelaar, Jonasson, and Besner 1977; Hahn, Chater, and Richardson 2003). A crude but often effective estimate for the degree of lexical support is the NEIGHBORHOOD DENSITY, defined as the number of existing words that differ from the nonword by a single modification (Luce 1986). Under this metric, a non-word like *plake* [pleɪk] would receive strong lexical support, since there are many similar existing words (*plate*, *lake*, *break*, etc.), while the non-word *plofe* [ploʊf] receives very little support (*loaf*).

Bailey and Hahn (2001) point out that limiting similarity to a single modification is generally too restrictive for drawing distinctions among non-words, since the majority of non-words have zero neighbors—a problem that is only exacerbated for sets of non-words with unattested clusters. Bailey and Hahn propose to overcome this problem by relaxing the notion of neighborhood to take into consideration words with smaller degrees of overlap. Intuitively, this has the potential to allow words like *bwadd* [bwæd] to receive support not just from single modification neighbors (of which it has just one: *bad* [bæd]) but also from more distant words, including those with #C*w* clusters (*quack*, *swag*, *quid*, etc.). As noted above, there are many more attested #C*w* clusters than #C*n* clusters in English (*#tw*, *#dw*, *#θw*, *#sw*, *#kw*, *#gw*[6] vs. *#sn*), raising the possibility that *#bw*-initial words might receive higher ratings solely on the basis of the fact that they overlap more with existing words.

The refinement of the traditional neighborhood metric that Bailey and Hahn propose is the Generalized Neighborhood Model (GNM), an adaptation of Nosofsky's (1986) Generalized Context Model. The Generalized Context Model is a similarity-based exemplar model that classifies new items according to their perceptual similarity to existing exemplars

(2) Generalized Context Model (GCM):

[6] And marginally *#mw*, *#nw*, *#fw*, *#vw*, all in French loanwords (*moire*, *noir*, *foie gras*, *voir dire*).

Probability of assigning novel item *i* to class of items C $=$

$$\frac{\sum_{c \in C} \text{Similarity(i, c)}}{\sum_{C'} \sum_{c \in C'} \text{Similarity(i, c)}}$$

The premise of the Generalized Neighborhood Model is that speakers assess the acceptability of a novel item by considering the probability of classifying it as a member of the set of English words. Since participants are not asked to assess the probability that the word is English as opposed to some other language, competition from other classes (the denominator) is irrelevant.

(3) Generalized Neighborhood Model (GNM):

Probability(novel word) $\propto \sum$ Frequency-weighted similarity(novel word,existing words)

In this model, the degree of support that a novel item receives from a set of existing items is not based on how many items meet a particular similarity threshold, but rather depends on the cumulative degree of similarity to all existing items. As an approximation of the set of existing words, we take the set of word forms that occur with non-zero frequency in CELEX (Baayen, Piepenbrock, and van Rijn 1993). The degree of similarity between a novel word and an existing word is taken to be a function of the transformations that are required to turn one into the other (Hahn, Chater, and Richardson 2003; Hahn and Bailey 2005). This requires finding the optimal alignment between the two words, such that phonetically similar segments are aligned with one another and as few segments are left unaligned as possible. The similarity of segments is estimated using the natural class based model of segmental similarity proposed by Pierrehumbert (1993) and defended in Frisch (1996) and Frisch, Pierrehumbert, and Broe (2004). The cost of unmatched segments (requiring insertion or deletion) is a free parameter of the model. Using these values, the minimum string edit (=Levenshtein) distance between the novel word and the existing word is calculated (Kruskal 1983/1999, chap. 1; Jurafsky and Martin 2000, §5.6). This transformational distance is then converted into a perceptual similarity score, as in (4):

(4) Similarity of two words (w,x) $= e^{(-d_{w,x}/s)^P}$, where

- $d_{w,x}$ = minimum string edit distance (w, x)
- $e \approx 2.71828$
- s, P = parameters of the model, which determine the size and nature of the influence of very similar neighbors; see Nosofsky (1986) for details

The similarity values for each existing word is then weighted according to a function of its token frequency, and the weighted similarities of all existing words are summed to yield an overall measure of support from the lexicon. For further details of the GNM, see Bailey and Hahn (2001). In the simulations reported here, an insertion/deletion cost of .7 was used, an s value of 0.1739, and a P value of 1.

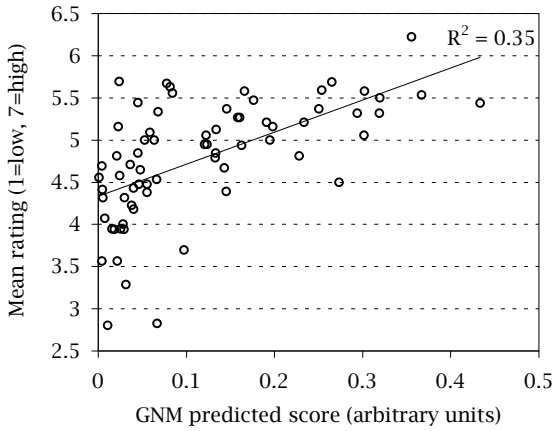


Figure 3: Benchmark performance of the GNM on 70 random filler items

Before we can evaluate whether the GNM does a good job in predicting differences among novel clusters, we must first assess how well the model does in capturing gradient acceptability judgments more generally. The rationale for this is the following: suppose that the model did very well in predicting differences among novel clusters, but did poorly in modeling arbitrary other differences (such as *plake* vs. *plofe*). In this case, we could hardly claim to have produced a general purpose similarity-based explanation of how speakers generalize to novel strings. Conversely, suppose that the model did badly on clusters, but also did badly in general. In this case, we might conclude that it is simply a bad model of lexical effects, or worse, a bad implementation of a reasonable model; in either case, we will have learned nothing. It is only possible to make inferences about the viability of a similarity-based account given a prior expectation that the model should do well based on its performance more generally when cluster well-formedness is not at stake.

To calculate benchmark performance of the GNM implementation, I used the model to derive predicted scores for the 70 randomly chosen filler items. The results are shown in Figure 3. As with other data sets, it appears that the GNM has difficulty distinguishing among relatively ill-formed items (see Albright, in prep., for discussion). Nevertheless, there is a reasonably good overall correlation between model and ratings (Pearson's $r(68) = .590$, $p < .0001$). We can take this as an indication that the GNM provides a generally decent model of gradient differences among words based on their degree of overlap with existing words.

We now turn to the predictions of the GNM for novel clusters. The GNM was used to derive predictions for the 40 non-words with {*p*,*b*}-initial clusters. As with the filler items, the correlation for these items was reasonably high ($r(38) = .636$, $p < .0001$). If we look in detail at the results in Figure 3, however, we see that this overall level of performance comes mainly from the fact that the model successfully predicts high scores for a handful of good cluster types (some *#pl-* and *#bl-*initial items), but fails to distinguish the remaining items in any meaningful way (the group of items on the left-hand side of the plot). The fact that it fails to predict any systematic preference for some clusters over others can be seen from the fact that onset clusters of all types are assigned low scores (*#pl, #bl, #br, #pn, #pw,* etc.). When the model's predictions are grouped by cluster, we

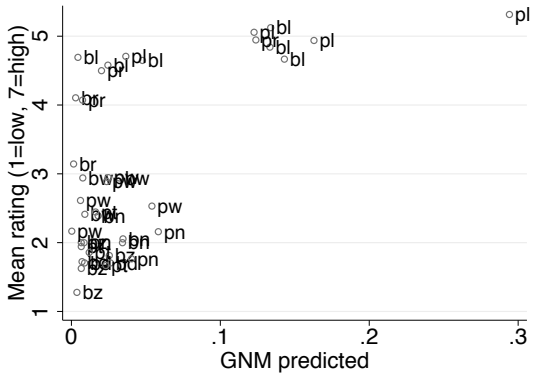


Figure 4: GNM performance on {*p*,*b*}-initial onset clusters

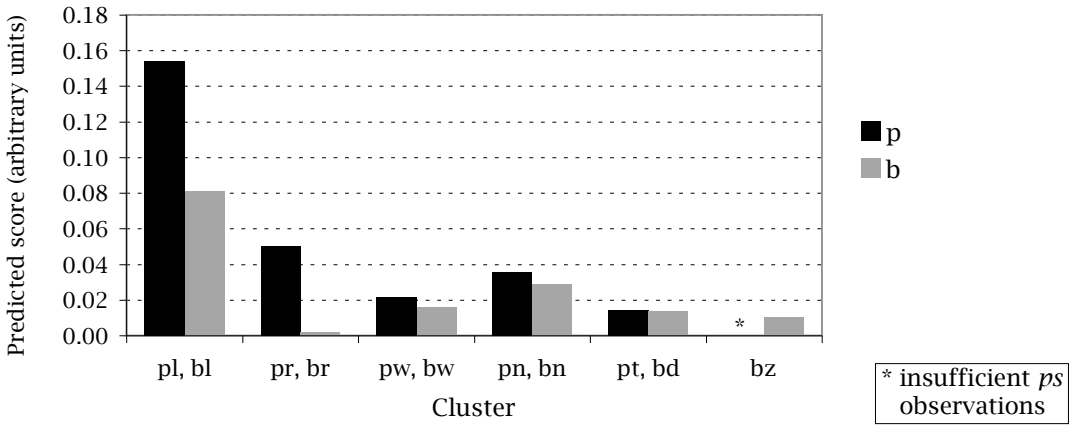


Figure 5: GNM performance grouped by cluster

see that apart from a preference for *#pl* and *#bl*-initial clusters, the model captures almost nothing of the effect observed in Figure 1 above. In particular, no consistent preference is predicted for *bw* ≻ *bn* ≻ *bz*, *bd*, contrary to what was observed in the experimental ratings. We conclude that lexical similarity alone is not a sufficient mechanism to explain English speakers' preference for some novel onset clusters over others.

The unsuitability of neighborhood similarity as an explanation for cluster preferences can also be seen in another way. In fact, relying on lexical similarity predicts that there could be substantial word-by-word differences, depending on what nearby words happen to exist. In some cases, this could even lead to preference reversals: *bwick* ≻ *bnick* due to similarity with *brick*, but *bneese* ≻ *bweese* due to similarity with *niece*, *lease*. More extreme examples would be non-words like [bnɛkfəst] or [bdənænə], which are predicted to be better than words like [bwæd] by virtue of the fact that they are very similar to existing words (*breakfast*, *banana*). Although the batch of non-words considered here cannot test this hypothesis (since they were selected to avoid close similarity to existing words), my own intuitions, for what they are worth, suggest that this prediction is

incorrect: [bwæd] is much more acceptable as a potential English word than [bdənænə].

The upshot of this section is that although a similarity-based analogical model does reasonably well at modeling gradient acceptability of arbitrary chosen nonce words containing attested sequences, it cannot simulate the observed preferences among unattested clusters. This failure is due primarily to the fact that the model is too sensitive to accidents of attestation throughout the entire word, and has no built in mechanism that would encourage it to focus on the initial cluster.

3.2 A perceptual confusability account

The conclusion of the previous section naturally leads to a related but different hypothesis: perhaps when speakers are asked to evaluate words that contain novel clusters, their attention is drawn mainly to the cluster itself (Sendlmeier 1987), and they judge its acceptability based on how similar it sounds to the nearest existing cluster. This is reminiscent of the idea of *perceptual assimilation*, which claims that listeners are perceptually biased towards legal sequences and may misperceive illegal sequences as similar existing ones. Perceptual assimilation has been viewed variably as a weak bias that affects ambiguous tokens (Pitt 1998; Moreton 2002), or as a strong perceptual constraint that can completely prevent listeners from hearing illegal sequences (Dupoux, Kakehi, Hirose, Pallier, and Mehler 1999). In the present case, the strongest possible version of this approach, in which linguistic signals are automatically mapped onto well-formed native structures by the perceptual system without any higher level processing, could not be right, since subjects were generally able to hear and repeat the "illegal" onset clusters at 90% accuracy rates even for the least acceptable clusters *#bz*, *#bd*. Nonetheless, it seems possible that even if subjects were able to hear the novel clusters accurately, the same mechanism that maps ill-formed sequences onto the perceptually closest legal sequence could be used to calculate a distance value between the incoming signal and the nearest attested sequence. This raises the question of whether the observed preferences could be explained by similarity of illict C_2's to attested or legal C_2's.

In order to answer this question, we need a model of perceived similarity (or confusability) between clusters. In theory, the nearest attested cluster could be one that differs either in C_1 or C_2 (or both). For *#pw*, *#bw* clusters, the existence of #C*w* provides highly confusable nearby analogs (*#kw*, *#gw* in particular), leading to a high predicted score. For less sonorous C_2's, on the other hand, there are no easily confusable clusters that differ only in C_1; *#bn* is scarcely confusable with *#sn*, and there are no attested #Cz, #C*d* clusters. In these cases, the closest attested cluster is likely to differ in C_2. The exact confusability values would naturally differ depending on the context, but as a rough estimate, I give in Table 1 the context-free similarity values calculated using the natural class method of Frisch, Pierrehumbert, and Broe (2004).

The values show something of the desired result: [n] is quite similar to [l] and [r], while [d] and [z] are less so. This suggests that perhaps *#bn* is rather well supported by its similarity to *#bl*, while *#bd* and *#bz* are more of a stretch. We note that support for [w] by analogy to [l] and [r] is quite weak; however, as described in the preceding paragraph, the relative goodness of *#bw* is plausibly due to its confusability with *#gw*, *#dw*, and not its similarity to *#br*, *#bl*.

Problems arise when we look beyond this set of segments, however, and consider contextual

Table 1: Estimated similarity values (higher = more similar, arbitrary units)

	d	z	n	l	r	w
d	1.000	.351	.406	.226	.200	.068
z	.351	1.000	.206	.259	.226	.075
n	.406	.206	1.000	.526	.435	.118
l	.226	.259	.526	1.000	.625	.148
r	.200	.226	.435	.625	1.000	.296
w	.068	.075	.118	.148	.296	1.000

allophones in C_2 position in a broader array of clusters. In particular, the devoicing observed in clusters like *#pl*, *#pr* creates highly fricated elements in C_2 position ([p̥l], [p̥r]), to the extent that /tr/ and /tʃ/ are practically indistinguishable for many speakers. This makes the interesting prediction that clusters like *#pʃ*, *#kʃ* should be very acceptable, because of their close similarity to *#pr̥*, *#kr̥*. There is, unfortunately, only a very small amount of data on this point: one *#pʃ*-initial non-word was included among the 100 filler items not included in either the analysis of clusters or benchmarking data (*pshuzz* [pʃʌz]). This item elicited very low ratings, almost exactly identical to the mean of *#bz*-initial words (1.6 on the scale of 1 (low) to 7 (high))—in spite of the fact that *#pʃ* is marginally attested in the exclamation *pshaw*. In other words,*#pʃ* is evidently judged to be very unacceptable, in spite of its close perceptual similarity to *#pr*. I take this to indicate that what is at stake in judging the acceptability of novel clusters is not some measure of raw similarity to attested clusters.

The problem with overall phonetic similarity is that it depends on multiple acoustic dimensions. By hypothesis, the licensing of consonants in C_1 position of clusters is a function of only certain dimensions—namely, the ones that affect the ability of C_2 to support cues to C_1. The similarity of [n] to [l] and [r] is exactly the relevant kind of similarity: sufficiently strong voicing amplitude and clear formant structure to permit the burst and transitions from C_1 to be clearly perceived. Thus, I am not ruling out an account in which the preference for *#bw*, *#bn* is tied to their acoustic similarity to *#bl*, *#br*, or even *#ba*. Crucially, however, in order to focus on this one relevant dimension of similarity, the model must have prior knowledge about perceptibility of contrasts, and a bias for onset clusters that maximize the perceptibility of C_1. A model of exactly this sort is discussed and endorsed in section 5. It is quite different from a model that rests solely on "innocent" perceptual assimilation. As we have seen, the latter type of model does not appear to be sufficient to model the observed preferences in onset clusters.

3.3 Local summary

To summarize the results of this section, we have seen that two simple models of how novel clusters are compared to existing clusters—in one case analogically, and in the other perceptually—cannot predict the observed preferences for *#bw* $\succ$ *#bn* $\succ$ *#bd*, *#bz*. Although this conclusion by no means proves that a grammatical account is indispensable, it does strengthen the argument by eliminating

some very reasonable alternative lines of explanation. In the next section, I turn to a different side of the problem: assuming that a grammatical account is needed, is it in fact unlearnable given the data of English?

4 An inductive model of grammatical learning

Even without direct evidence about *#bn*, *#bd*, English learners do get plenty of evidence about stop+sonorant (or even stop+consonant) sequences from existing sequences like *bl*, *br*, *sn*. The models in section 3 attempted to make use of this evidence based on raw similarity. A different intuition, in line with work in theoretical phonology, is that existing clusters provide evidence about natural classes of segments that can co-occur in a particular position. By comparing feature values in attested clusters, learners might discover that they can ignore or recombine particular feature values to predict the possibility of unseen cluster types, as in (5).

(5) Generalization based on natural classes

a. Interpolation: *#br*, *#sn* $\Rightarrow$ $\begin{bmatrix} -\text{syllabic} \\ -\text{sonorant} \end{bmatrix}\begin{bmatrix} -\text{syllabic} \\ +\text{sonorant} \end{bmatrix}$

b. Extrapolation: *#br*, *#bl* $\Rightarrow$ $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{voice} \\ +\text{labial} \end{bmatrix}\begin{bmatrix} -\text{syllabic} \\ +\text{sonorant} \end{bmatrix}$

In this section, I sketch a model that is designed to learn constraints on possible two-segment sequences stated in terms of natural classes, and to evaluate the amount of support that they get based on statistical properties of the linguistic data.

4.1 A model for discovering and evaluating natural classes

It is a firmly held tenet of generative phonology that speakers can generalize patterns based on knowledge of feature combinations. An example of this, almost as famous as the 'blick' test, is the 'Bach' test (Halle 1978): for the subset of English speakers who can produce the voiceless velar fricative [x], the plural of 'Bach' is [baxs] with a voiceless [s], not *[baxz] or *[baxəz]. Although the segment [x] is systematically absent from the training data of English, this generalization is supported by the fact that all featurally equivalent segments (the voiceless non-strident sounds: {p, t, k, f, θ}) regularly pluralize with [s]. In order to learn this distribution, speakers must have a way of comparing words that pattern alike and extract the feature values that may be relevant to their behavior.

One model for how learners abstract away from individual segments to natural classes is the MINIMAL GENERALIZATION approach (Albright and Hayes 2002, 2003, 2006). Under this approach, learners compare sequences of sounds pairwise, aligning them and extracting what feature

values they have in common. Shared feature values are retained and encoded in a more abstract rule/constraint, and unshared values are eliminated ((6)).

(6) Abstracting over strings by minimal generalization

	b	l	u
+	b	r	u
→	b	[+consonantal, +sonorant, −nasal]	u
+	g	r	u
→	[−sonorant, −continuant, +voice]	[+consonantal, +sonorant, −nasal]	u

Although the mechanism for abstracting over segments to larger classes is intuitive, it is not so obvious which segments should be compared with which. In the example in (6), the comparison is quite natural because [bl] and [gr] are very similar and define a space of combinations that could reasonably be expected to pattern together ([bl], [gl], [br], [gr]). Not all comparisons are so informative, however. When dissimilar clusters are compared, the resulting inference can be an extremely broad:

(7) A less informative comparison

	b	l	a
+	s	p	a
→	[+consonantal, −nasal, −lateral]	[+consonantal, −nasal, −strident]	a

In some cases, broad inferences are good; they allow the learner to discard irrelevant details of the training examples and extract the more general pattern. In this case, the resulting pattern is one in which combinations of non-nasal, non-strident consonants may co-occur—a pattern which is indeed very well attested in English (*bl*, *gr*, *sp*, etc.). Unfortunately, this also leads to the potentially fatal prediction that clusters like *#bd* should also be very acceptable, since they too fit the pattern in (7). The challenge is to find a way to generalize over natural classes such that initial *#bl* and *#br* provide moderate support for *#bn*, even though it is outside the feature space that they define, while comparisons like *#bl* and *#sp* should not support generalization to *#bd* even though it is within the space that they define.

The solution is to relax the minimal generalization assumption, while at the same time penalizing sweeping generalizations that go too far beyond the set of attested examples. Intuitively, the problem with the inference in (7) is that the examples support it only very "spottily". If [+consonantal, −nasal, −lateral] [+consonantal, −nasal, −strident] sequences are really allowed in English, then we should expect to find a large and diverse set of examples supporting this inference. In fact, the attested examples that support this generalization are grouped into a few particular subgroups (obstruent+liquid

clusters, sC clusters, etc.). The broader generalization fails to explain why only certain subtypes are actually attested, and therefore does not accurately characterize the distribution of attested examples. In order to penalize overly broad generalizations, we must take into account not only the frequency with which a particular combination of natural classes occurs in the data, but also how strongly the combination of natural classes leads us to expect the particular combinations of segments that we observe. The feature combination $\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix}\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix}$ is quite frequent, but does not strongly predict any particular combination of surface consonants. The feature combination $\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix}\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix}$ on the other hand is not as frequent, but because it is so specific it strongly leads us to expect to encounter surface sequences like [br] or [gl]. For further discussion of this point, and the implementation of the penalty, see Albright (in prep.).[7]

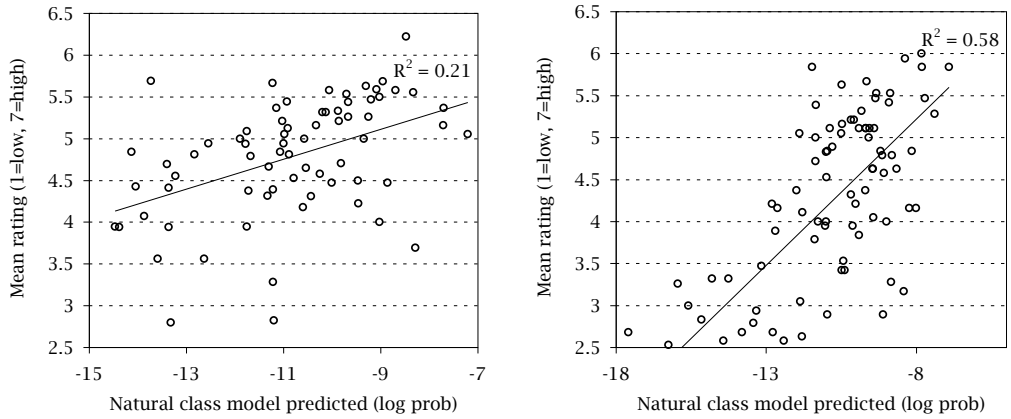
The output of this model is a list of sequences of natural classes with scores attached to them. These scores reflect a combination of the frequency of each sequence in the data, and the ability of the sequence of classes to predict the specific training items. Novel words are evaluated by parsing them into combinations of natural classes, attempting to characterize each two-segment substring using the most probable possible combination of natural classes. For words that consist of attested bigrams, the best way to parse them is typically into the specific combinations of segments that they contain. For words with unattested sequences, the model must seek a more general way to parse the novel substring, in a way that groups it with attested substrings.

4.2 Testing the model, part 1: Benchmarking

The model outlined in the previous section provides an account of how learners encode statistical properties of their language and generalize them (gradiently) to novel items. Once again, it is important to ask how well this model performs on arbitrarily selected words consisting of attested combinations, before asking the more interesting question of what it predicts for unattested combinations. In order to test this, the model was trained on an input file containing all of the word forms found in CELEX with frequency > 0, and then used to derive predictions for the batch of 70 benchmark items used above.

The results are shown in Figure 6a. The model achieves a moderate correlation to subjects' ratings ($r(68) = .454$, $p < .0001$). This is not as good a fit numerically as the GNM (Figure 3), but we see that the model's predictions are also more consistent throughout the entire range. In defense of the claim that this model achieves reasonably good baseline performance on arbitrary batches of non-words, its predictions for a different set of nonwords are shown in Figure 6b ($r(90) = .759$, $p < .0001$) (Albright and Hayes 2003). Although for the novel clusters we might expect performance closer to that in Figure 6a (since the data is from the same experiment), the comparison data is provided to show that the result is not accidental; although the exact fit varies, the model does seem to provide fairly good predictions for novel items across different sets of data.

[7] A similar problem is discussed in Albright and Hayes 2000, and a different solution is suggested there.



(a) Performance on 70 nonwords from §2

(b) Performance on 92 non-words from Albright and Hayes (2003)

Figure 6: Performance of the natural classes model on attested sequences

4.3 Testing the model, part 2: Scholes (1966) onset data

As mentioned in the introduction, the idea that preferences among unattested clusters might be predictable based on the set of existing clusters has already been confirmed to a certain extent by the results of Hayes and Wilson (in press) in modeling data from Scholes (1966). In the Scholes study, words with attested and novel onset clusters were presented auditorily to 33 seventh graders, embedded in words with attested rhymes: e.g., *skeep* [skiːp], *mlung* [mlʌŋ], *flurk* [flr̩k], *zhpale* [ʒpeɪl]. Subjects gave binary yes/no decisions about whether the words were possible words of English.

(8) Scholes onset clusters

a. Attested clusters

#pl, #kl, #bl, #gl, #fl, #sl
#pr, #tr, #kr, #br, #dr, #gr, #fr, #ʃr
#sp, #st, #sk, #sf, #sm, #sn, #ʃn

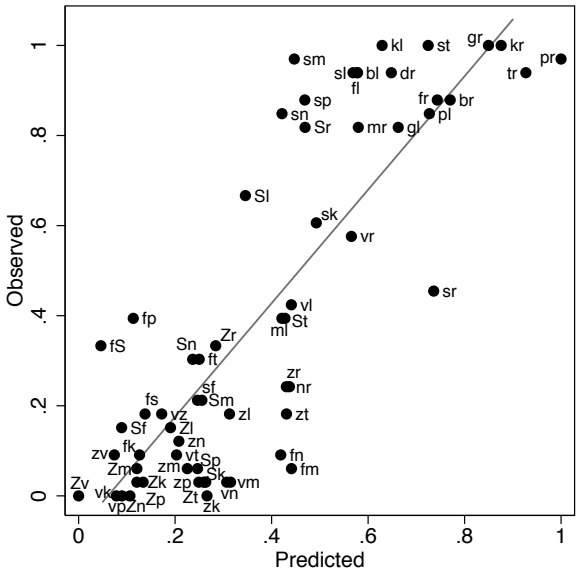


Figure 7: Performance of the natural class model for onset cluster data (Scholes 1966)

b. Unattested clusters

#tl, #dl, #ml, #nl, #vl, #zl, #ʃl, #ʒl, #mr, #nr, #vr, #sr, #zr, #ʒr,
#fp, #vp, #zp, #ʃp, #ʒp, #ft, #vt, #zt, #ʃt, #ʒt, # fk, #vk, #zk, #ʃk, #ʒk,
#fm, #vm, #zm, #ʃm, #ʒm, #fn, #vn, #zn, #ʒn,
#fs, #vz, #fʃ, #vʒ, #zv, #ʃf, #ʒv

Hayes and Wilson test the ability of their inductive model of constraint learning to predict the proportion of "yes" responses to each cluster by training the model on a corpus of attested English onsets. The corpus consisted of word-initial onsets found in the CMU Pronouncing Dictionary,[8] filtered to remove onsets that they felt to be "exotic" or non-native (e.g., *#sf*, *#zw*, *#ʃn*). The model uses the input set to learn a set of constraints, which are then numerically weighted in such a way that the grammar is able to assign gradient well-formedness scores to attested and unattested onsets. Hayes and Wilson find an impressively close fit between the models' predicted scores and the proportion of "yes" responses by Scholes' subjects ($r(60) = 0.946$), proving that it is in principle possible to learn differences among unattested clusters based on positive evidence from attested clusters.

As a point of comparison, the model described in section 4.1 was trained on the "filtered" CMU dictionary corpus and used to derive predicted scores for the set of onset clusters tested by Scholes (1966). As in the Hayes and Wilson study, the predictions of the model were transformed according to a function of the form $y = x^{1/k}$ (here, k=2.01) and rescaled to the range (0,1) in

[8] http://www.speech.cs.cmu.edu/

order to facilitate comparison with the observed proportions of "yes" responses.[9] The result can be seen in Figure 7, which shows a reasonably good fit between the model's predictions and subject responses (Pearson $r(60) = .830$; Spearman $r(60) = .793$). Although the fit is decent, it must be noted that this model does not achieve quite as good a fit as what Hayes and Wilson report for the models that they compare, which range from r=.833 to .946 (Spearman $r = .757$ to .889). This raises the possibility that the current model is a poor starting point for any further evaluation of whether more fine-grained distinctions (such as *#bn* $\succ$ *#bd*) are learnable. If we look more closely at the results, however, there is reason to think that the model should not be dismissed outright.

A potentially redeeming virtue of the current model is that it attempts to differentiate both among pairs of attested and also pairs of unattested items. Comparing the distribution of predictions in Figure 7 against the predictions of the Hayes and Wilson model (their number (17)), it is clear that the major difference is that the current model produces a wide spread of predicted values for attested clusters (upper right portion of the plot), while the Hayes and Wilson model assigns nearly all attested clusters a score of 1. In many cases, the willingness of the current model to assign less than perfect scores to attested clusters means that it seriously underestimates their goodness. This effect is particularly noticeable for *#sC* clusters (*#sm*, *#sn*, *#sl*, *#sp*), and mirrors the relative rarity of these clusters in the training corpus. On the other hand, it is precisely the ability to differentiate among attested clusters that allows the model to do well on benchmarking data involving arbitrary well-formed sequences (Figure 6). The Hayes and Wilson model achieves a good numerical fit by assigning all of these clusters the same maximum score, since Scholes' subjects also tended to agreement nearly unanimously that words with attested clusters were possible words of English. It seems possible that at the high end of acceptability, this task fails to tap into fine-grained intuitions about relative well-formedness, and that a more sensitive ratings task would reveal systematic preferences for some existing clusters over others.[10] Conversely, some of the models considered by Hayes and Wilson fail to differentiate among unattested sequences. Although comparable graphs are not given, a "classic" *n*-gram model defined over phones would assign scores of 0 to all words with unattested sequences, and analogical similarity-based models tend to have a large cluster of predicted values at or near zero as well (for examples, see Albright, in prep.). The current model has the virtue of making gradient predictions both for attested and also unattested sequences. Practically speaking, this has the consequence that the model's predictions have a less skewed or bimodal distribution, making it difficult to carry out exact numerical comparisons of fits across different modeling results. More important, even if the details of these gradient predictions are not always perfect (viz. the *#sC* clusters in Figure 7), they permit the model to do reasonably well across a range of applications.

We must bear in mind when comparing performance on different tasks that the results in this section differ from those in previous sections in that (following Hayes and Wilson) the model was trained and tested specifically on onset clusters rather than entire words. This may be seen as a

[9] Given that subjects' responses tend to cluster at the lower and upper ends of the scale, it seems natural to use a higher-order polynomial fit between the model's scores and observed values. This option was not pursued here, both in order to facilitate comparison with the Hayes and Wilson results, and also because preliminary analysis revealed that more complex models would not yield a substantially better fit.

[10] Indeed, the results of Figure 4.1 do show such preferences, although it is impossible to attribute any item-by-item differences specifically to clusters vs. the vowel-consonant combinations that they contain.

rather artificial restriction, since Scholes' subjects were necessarily presented with entire words (*skeep*, *mlung*, *flurk*, etc.), and not clusters in isolation. Hayes and Wilson plausibly suggest that the blandness of the rhymes that Scholes selected may make their contribution negligent to the overall results. This explanation is ambiguous, however: the rhymes may be ordinary in such a way that they would be assigned the same values by gradient phonotactic models, or they may have been bland in such a way that subjects ignored them in spite of potential numerical differences between them. In order to tease apart these hypotheses, I trained and tested the model in several different conditions: in the first condition, the model was training on the Hayes and Wilson "filtered" onsets corpus from the CMU Pronouncing Dictionary and novel words were assigned scores solely according to their onsets, as reported above. In the second condition, the model was training on the full set of lemmas in CELEX with frequency > 0, but novel items were assigned scores based just on the likelihood of the onset cluster (i.e., the bigrams #C_1 and C_1C_2). Next, the model was trained on the set of CELEX lemmas, but scores were assigned based on the product of the scores for the onset and rhyme (i.e., ignoring the C_2V transition to simulate onset-rhyme independence). Finally, the model was trained on the set of CELEX lemmas and novel words were assigned scores according to their complete set of bigrams, as in section 4.2 above. The results in (9) clearly show that the rhymes in Scholes' test items were not statistically equivalent, but rather, are predicted to make a substantial difference in ratings (seen in the difference in performance between the three modes of testing the CELEX model). However, the best model of the responses is one in which the rhymes are ignored completely.

(9) Performance of the model on Scholes cluster data, under different training conditions

Training set	Score based on	Pearson *r*
CMU Dict. (filtered), onsets only	#C_1C_2	.830
CELEX lemmas	#C_1C_2	.713
CELEX lemmas	Onsets, rhymes separate	.699
CELEX lemmas	Whole string combined	.503

These results support the idea that by including just a few rhymes that were repeated over and over, Scholes cued subjects to ignore the rhymes and focus their attention on the onsets, which were unusual and varied.[11] This is also reminiscent of the finding of Sendlmeier (1987) that at least for relatively simple words, naive listeners tend to focus on particular salient features of novel items. Certainly, not all experimental settings encourage such selective attention to a particular region of the word, however. For example, the results in Figure 6 require evaluation of the entire word, and Coleman and Pierrehumbert (1997) discuss cases in which very improbable sequences in one portion of the word are offset by very probable sequences elsewhere. One might plausibly hypothesize that in tasks with a large variety of rhymes—including the experiment described in section 2—subjects' decisions should not be based so exclusively on any single part in the word. Nonetheless, in the absence of a clear-cut set of principles for predicting ahead of time what the most appropriate model of acceptability ratings should be, the most generous strategy is to try both

[11] The difference between the CMU and CELEX training sets is also greater here than what Hayes and Wilson find, but this may be due to the fact that in the present case, training included words with and without onsets and the model's attention was focused on onsets only in testing. It is also possible that the choice of CELEX word forms vs. lemmas is important.

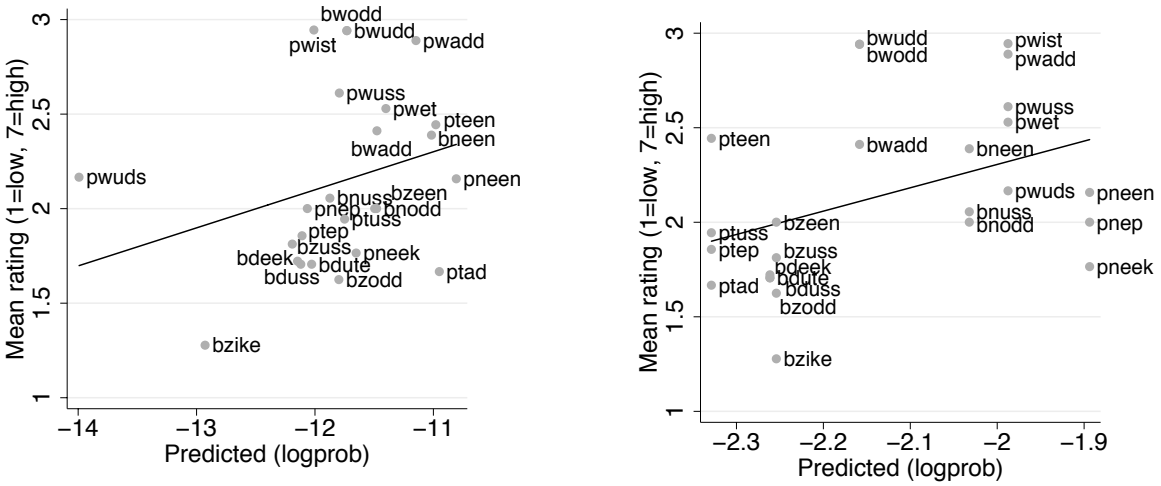
types of model (whole word, vs. onset only). If the model succeeds at predicting the observed distinctions between clusters under either mode of testing, we will tentatively conclude that the distinctions are learnable, with the provision that we must eventually be able to explain when subjects focus on a particular aspect of novel items and when they evaluate them more holistically.

In sum, the results of this section show that the proposed model is able to model some preferences among onset clusters, provided that we are willing to assume that at least in some cases, subjects behavior is based solely on the clusters in question and not on any other aspect of the word that was presented to them. This provides the necessary backdrop for comparing the model's performance on the specific clusters of interest—namely, those with differing sonority profiles (*#bw* vs. *#bn* vs. *#bz*, *#bd*)

4.4 Testing the model 3: sonority sequencing

We turn finally to the question of whether the model is able to generalize appropriately to novel onset clusters like *#bw*, *#bn*, *#bz*, and *#bd*. Based on the results of the preceding section, the model was trained in two different ways: once using the corpus of lemmas from CELEX (as in section 4.2), and once using the "filtered" set of onsets from the CMU pronouncing dictionary (as in Hayes and Wilson). The model was then tested on the set of words/onsets from the experiment in section 2. As Figure 8 shows, neither testing scenario provides a particularly good model of subjects' ratings, though the onsets-only model performs substantially better (whole word: $r(38) = .285$, $p = .08$; onsets only: $r(38) = .405$, $p < .01$). The greater success of the onsets-only model is perhaps a bit suspect in this case, since unlike the Scholes study, the experimental design involved a significant number of filler items (170) with a wide variety of rhymes, making predominantly onset-based decisions seem unlikely. Furthermore, inspection of the plot in Figure 8b reveals that subjects did not appear to be ignoring the rhymes, since although pairs involving controlled rhymes did come out quite similar (e.g., *bdeen*, *bdute*, *bduss* all at more or less the same vertical height), in cases where the rhyme was considerably less likely (e.g., [ʌdz] in *pwuds*, [aɪk] in *bzike*), the novel items received correspondingly lower ratings. Nonetheless, in order to provide maximum benefit of the doubt, we will take the results from the onsets-only training (Figure 8b) as indicative of the model's ability to learn distinctions among this set of onset clusters.

Numerically, the correlation between the model's predicted scores and the mean observed ratings is significant. This could be taken as evidence that the model is at least somewhat able to learn differences among this set of onset clusters, even without an explicit prior notion of sonority. Figure 9 shows the model's predictions grouped by onset cluster; some preferences, such as *#br* $\succ$ *#bn* and *#bn* $\succ$ *#bz, #bd* are indeed successfully predicted. At the same time, the model completely fails to predict other preferences, such as *#bl* $\succ$ *#pn* or *#bw* $\succ$ *#bn*. Whereas subjects expressed clear preferences for some combinations over others (reflected in the fact that the points for *#pw*, *#bw* are high in the plot, followed vertically by *#pn*, *#bn*, then *#pt*, *#bz*, and *#bd*) neither model recapitulates this systematically in the horizontal dimension. This failure is most evident in Figure 9, which shows the incorrectly reversed prediction *#pn*, *#bn* $\succ$ *#pw*, *#bn*, and the fact that *#bl* has approximately the same predicted value as *#pn*. I conclude that in spite of the significant overall correlation, the model has not successfully learned the set of observed preferences.



(a) Whole word score, CELEX training

(b) Onset cluster score, CMU dict. onsets training

Figure 8: Performance of the natural classes model on {*p*,*b*}-initial onset clusters

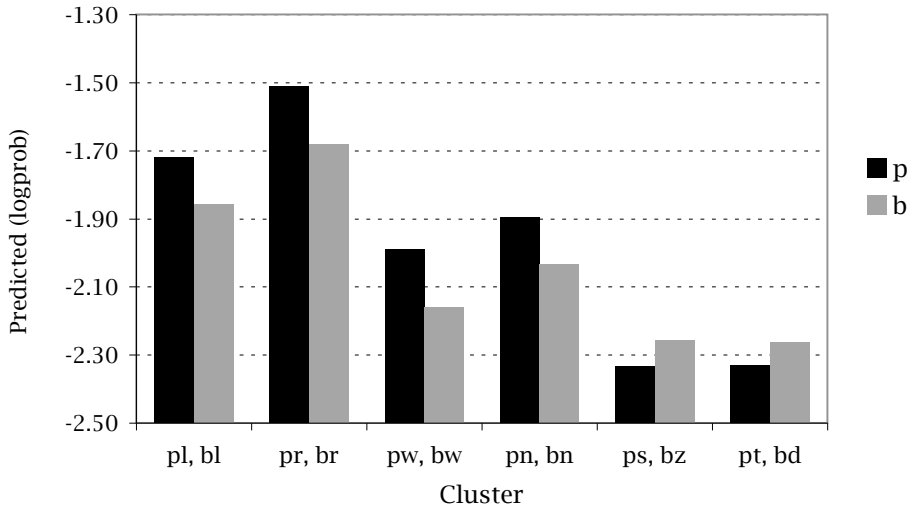


Figure 9: Performance of the natural classes model by cluster (onsets-only model)

Table 2: Antagonistic CC vs. CV transitional probabilities

	bC_2	C_2V
blV	high	low
bwV	low	very low
bnV	very low	low
bdV	extremely low	high

There are apparently several factors contributing to this failure. For the whole-word model, the somewhat correct predictions that could be made on the basis of onsets alone are counteracted by predictions based on the combination of onset and nucleus, nucleus and coda, and so on. Even if the model succeeded perfectly in predicting the local preference for *#bw* $\succ$ *#bn* $\succ$ *#bd*, there is a different and sometimes conflicting set of preferences for onsets: [d] is a much more common word onset than [n] or [w] (which very often find themselves in word-medial or coda position). These complicated and sometimes conflicting relations are shown schematically in Table 2. For an unbiased statistical learner, there is no reason to attach any greater significance to the cooccurrence of consonants in onset clusters than to any other cooccurrence of adjacent elements.

The onsets-only model avoids this problem by not considering the contribution of onset-nucleus or nucleus-rhyme combinations. However, it still suffers from a different problem concerning feature-based generalization. The set of features that is standardly assumed in phonological analyses makes it easy to group individual consonants with sets of existing C_2's. This is illustrated in (10), which shows that generalization to different C_2's can be based on local groupings with attested C_2's. As a result, there is no guarantee that implicational relations will fall out as an automatic consequence of feature-based extrapolation.

(10) Narrow groupings of potential C_2's

a. {*l*,*r*,*n*}: $\begin{bmatrix} +\text{sonorant} \\ +\text{coronal} \end{bmatrix}$

b. {*l*,*r*,*d*}[12]: $\begin{bmatrix} +\text{voice} \\ +\text{coronal} \\ -\text{nasal} \end{bmatrix}$

c. {*l*,*r*,*z*}: $\begin{bmatrix} +\text{voice} \\ +\text{anterior} \\ +\text{continuant} \end{bmatrix}$

Both of these issues point in the same direction: without an explicit bias to focus on C_1 and the sonority of the following element, no systematic sonority-based generalization emerges from the model. Thus, we are in a very similar situation as above: the model does quite well at generalizing

[12]In this particular case, privative nasality (i.e., no [−nasal] feature value) would succeed in making it difficult to characterize *d* without also including *n*. This local solution would not address the more global problem, however.

to novel words when gradient preferences among attested sequences are involved, but it is unable to generalize correctly to novel words with unattested sequences. The maneuvers explored here that are designed to help generalize beyond the set of attested clusters unfortunately tend to overpredict the goodness of more marked clusters, and do not mirror the preferences that native speakers of English show for certain combinations over others. This helps further the claim that the evidence needed to differentiate these clusters is not obvious from the data of English. Of course, it is always possible that a more sophisticated model might make more headway in predicting these preferences, but once again, at least one simple and appealing idea has proven insufficient.

5 Incorporating prior biases

This result joins a growing body of literature showing that speakers display phonological preferences that do not mirror lexical statistics in any obvious way. This puts us in the difficult and unenviable position of arguing from a negative result: is it truly impossible to model cluster preferences based solely on the statistics of English, or are these merely inadequate models? For this reason, any conclusions that we draw are necessarily provisional: the preference does not appear to be fully learnable, given any currently known learning procedure. The fact that this preference does not appear to be learnable using any of the techniques that have been tried so far does not prevent us from providing a formal model of the effect, however. In this section, I follow Wilson (2006) in showing that by simply adding a phonetically motivated bias to a statistical model, we can provide an overall model that closely mimics speaker preferences.

The preference for consonants to occur before more sonorous elements at the beginnings of syllables has been widely discussed in the phonological literature, and is often referred to as the SONORITY SEQUENCING PRINCIPLE (SSP). The experimental results in section 2 bear on just one subpart of the SSP: a preference for stops to be followed by more sonorous elements. This preference has a plausible phonetic motivation (Steriade 1997): in order to have its place and laryngeal features perceived accurately, a stop must have a clear burst and formant transitions for the stop closure must be readily apparent in the surrounding segments. Stop cues are jeopardized by segments with less clear formant structure (nasals, and even worse, obstruents), and by segments with formant targets of their own that would obscure the stop's transitions (e.g., [w] after [b], or [l] after [d]). These considerations favor putting stops before before segments that have greater voicing amplitude, clear formant structure, and a lack of interfering formant movement caused independently by C_2.[13]

For present purposes, I will treat the greater licensing of stops before more sonorous elements and the ban on antagonistic place combinations as two distinct constraints, though ultimately it may be preferable to state them as a unified condition on possible contrasts. For the model developed here, the constraint preferring that stops occur before sonorous elements is stated as a positive

[13]The more abstract form of the Sonority Sequencing Principle (a monotonic increase in sonority from syllable margin to nucleus, preferring steeper rises over shallow rises, shallow rises over level sequences, and level sequences over sonority reversals) covers a much broader range of segments, including the liquid+stop clusters discussed by Berent et al. (in press) (**lbif*). I focus here on the better understood conditions on stops in C_1 position, leaving aside for now the question of how and whether to unify this analysis with other cases traditionally covered by the SSP.

constraint demanding vowel-like voicing amplitude and formant structure, which is gradiently violated by consonants of decreasing sonority. In recognition of the fact that liquids are considerably better than nasals in their ability to carry formant transitions, they are given a (somewhat arbitrary) difference of 4 violations, as in (11a); the exact value does not matter, though a constraint defined in this phonetically sensible way was found to work better than a linearly decreasing scale. The ban on sequences like *pw* and *tl* was treated as binary.

(11) Constraints on occurrence of stops in C_1 position

a. Stop / __ maximally sonorous element

Violations:

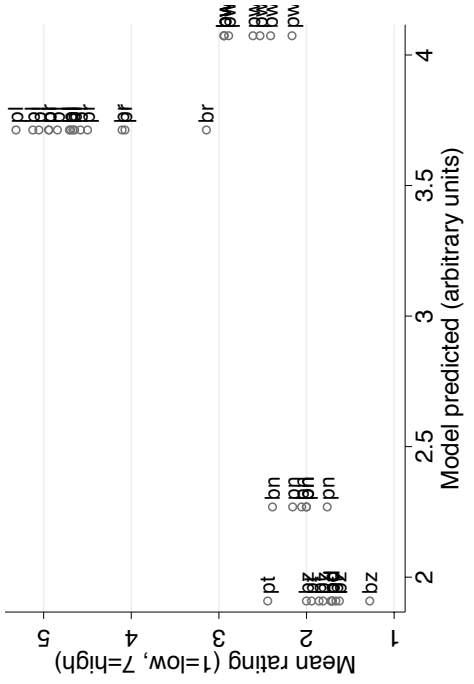
Stop+Glide	0
Stop+Liquid	1
Stop+Nasal	5
Stop+Obstruent	6

b. Specific place-related effects:

*{*pw*,*bw*}, *{*tl*,*dl*}

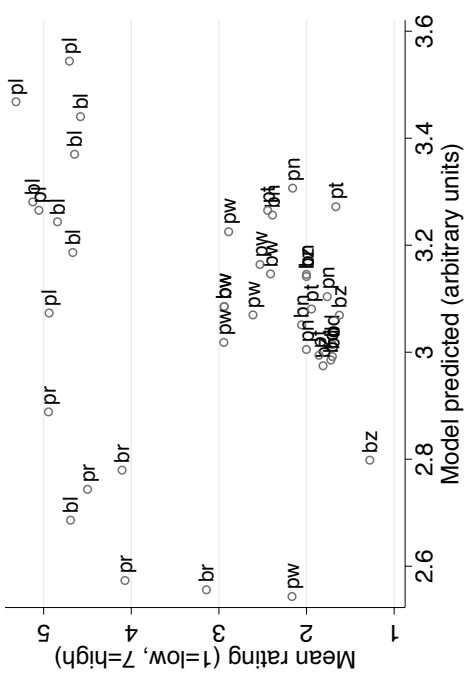
Rather than incorporating phonetic bias directly into the statistical learning procedure, as Wilson (2006) does, a post hoc method was adopted in order to assess the relative contribution of phonetic biases: the preferences of the inductive learning model from section 4 were combined with the constraints in (11) in a Generalized Linear Model. The predictions of the inductive model were taken to be the results of whole-word training and evaluation (Figure 8a), since the experimental task was one which should intuitively favor whole-word evaluation, and since inspection of the scatter plots revealed clear effects of the novel words' rhymes (see discussion above). The linear model combines three factors to assign score of a novel word: the statistical likelihood assigned by the inductive model, a numerical penalty for sonority violations ((11)), and a penalty for any ill-formed sequences like *pw* or *bw*. Each of these factors is assigned a weight, reflecting its contribution in determining overall well-formedness. Since we do not know ahead of time the importance that speakers place on the constraints in (11) compared to inductively learned statistical patterns, the relative weights were found post hoc by maximum likelihood optimization, attempting to find the best fit between the model's predictions and subjects' ratings. The relative contribution of different factors can be seen by observing the progression in Figure 10a–d. In the first step, we see that the inductive model alone is insufficient to predict judgments of novel onset clusters (recapitulating the result from section 4 above), and that the bias for stops to occur before segments with greater voicing amplitude and clearer formant structure is by itself also insufficent. In step 2, these two factors are combined ((11a)). This step significantly improves to the model, but leaves the relative dispreference for *pw*, *bw* unexplained (lower right-hand side of the plot). Finally, in step 3, the bias against *pw* and *bw* is added, yielding a model that is overall very accurate ($R^2 = 94\%$). All three factors contribute significantly to the final result, and could have been shown in any order; the choice to show the statistical model first and the incremental gain added by phonetic biases is a purely expository one.

The analysis presented here is hopelessly hand-crafted in several respects. Statistically learned and prior biases are treated as separate entities, rather than allowing bias to guide statistical learn-



(a) Step 1a: Inductive model alone ($r(38) = .182$, n.s.)

(b) Step 1b: Sonority bias alone



(c) Step 2: Incorporating preference for sonorous C_2 ($r(38) = .733$)

(d) Step 3: Incorporating bias against *pw*, *bw* ($r(38) = .971$)

Figure 10: Incremental addition of constraints to the full model

ing in a more integrated fashion (as Wilson (2006) does). Furthermore, the sonority scale, though guided by phonetic principles, has been hand coded. Finally, and perhaps most importantly, the relative magnitudes of statistical and phonetic preferences have been established post hoc by fitting to experimental data, providing us with no explanation as to how or why English speakers assign these relative weights to the factors involved. Nonetheless, I believe there is still some value to results like those in Figure 10d, since they show in a concrete and quantifiable way the failures of both the purely statistical and purely hand-coded approaches, and the gain that can be had from incorporating both universal and statistically learned preferences in a single model. A minor improvement to this post hoc model would be to attempt to estimate the coefficients for the constraints based on positive linguistic data; current work in linear constraint models of phonology suggest some strategies for this problem (Goldwater and Johnson 2003; Jaeger, to appear; Pater, in prep.). A more substantial improvement would be to incorporate bias directly into the statistical learning model, rather than imposing the bias externally as a separate module (Wilson 2006; Hayes and Wilson, to appear).

6 Conclusion

An increasing amount of attention has been devoted in the recent phonological literature to documenting cases in which speakers show preferences for some structures over others, in spite of the fact that both are equally robust or equally unattested in the input data of the native language. Such cases are of enormous interest because of their potential to reveal substantive phonological biases; however, actually proving that a universal grammatical bias is at play can be extremely difficult, and demands at minimum a good faith effort to attempt to infer the preference from linguistic data. The strategy adopted here is to pursue a range of conceptually simple and appealing data-driven approaches, showing that when implemented, none is sufficient to explain the observed preference for some unattested onset clusters over others. Although in principle this result could reflect implementational shortcomings rather than theoretical faults of the particular approaches, comparison with "benchmarking" results on arbitrary sets of non-words that lack fatal phonotactic violations reveals that the models are well suited to at least some tasks. This suggests (though does not prove) that a simple model based purely on statistical properties of the linguistic data is inadequate, just as one based purely on phonetic biases would be. Perhaps most tellingly, when both types of knowledge are combined, it is possible to construct an extremely accurate model of speaker preferences. It is hoped that such hand-crafted models can serve as both a challenge and a standard for success for less ad hoc models in the future, whether achieved through biased learning or more sophisticated data-driven means.

7 Appendix

Cluster	Word	Transcription	Mean rating
pl	plake	[pleɪk]	4.94
	pleen	[pliːn]	5.32
	pleek	[pliːk]	5.06
	plim	[plɪm]	4.71
	blute	[bluːt]	4.84
bl	blodd	[blad]	5.13
	bluss	[blʌs]	4.67
	blad	[blæd]	4.65
	blemp	[blɛmp]	4.69
	blig	[blɪg]	4.58
pr	prundge	[prʌnd͡ʒ]	4.94
	prupt	[prʌpt]	4.07
	presp	[prɛsp]	4.50
br	brelth	[brɛlθ]	3.14
	brenth	[brɛnθ]	4.11
pw	pwet	[pwɛt]	2.53
	pwist	[pwɪst]	2.94
	pwuss	[pwʌs]	2.61
	pwadd	[pwæd]	2.89
	pwuds	[pwʌdz]	2.17
bw	bwudd	[bwʌd]	2.94
	bwadd	[bwæd]	2.41
	bwodd	[bwad]	2.94
pn	pnep	[pnɛp]	2.00
	pneek	[pniːk]	1.76
	pneen	[pniːn]	2.16
bn	bnuss	[bnʌs]	2.06
	bneen	[bniːn]	2.39
	bnodd	[bnad]	2.00
bz	bzuss	[bzʌs]	1.81
	bzeen	[bziːn]	2.00
	bzike	[bzaɪk]	1.28
	bzodd	[bzad]	1.63
pt	pteen	[ptiːn]	2.44
	ptad	[ptæd]	1.67
	ptuss	[ptʌs]	1.94
	ptep	[ptɛp]	1.86
bd	bdute	[bduːt]	1.71
	bduss	[bdʌs]	1.71
	bdeek	[bdiːk]	1.72

References

Albright, A. Gradient phonological acceptability as a grammatical effect. MIT ms.

Albright, A. and B. Hayes (2000). Distributional encroachment and its consequences for morphological learning. In A. Albright and T. Cho (Eds.), *UCLA Working Papers in Linguistics 4 (Papers in Phonology 4)*, pp. 179–190.

Albright, A. and B. Hayes (2002). Modeling English past tense intuitions with minimal generalization. *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, 58–69.

Albright, A. and B. Hayes (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition 90*, 119–161.

Albright, A. and B. Hayes (2006). Modeling productivity with the Gradual Learning Algorithm: The problem of accidentally exceptionless generalizations. In G. Fanselow, C. Féry, R. Vogel, and M. Schlesewsky (Eds.), *Gradience in Grammar: Generative Perspectives*, pp. 185–204. Oxford University Press.

Baayen, R. H., R. Piepenbrock, and H. van Rijn (1993). *The CELEX lexical data base on CD-ROM*. Philadelphia, PA: Linguistic Data Consortium.

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*.

Chomsky, N. and M. Halle (1965). Some controversial questions in phonological theory. *Journal of Linguistics 1*, 97–138.

Cohen, J., B. MacWhinney, M. Flatt, and J. Provost (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers 25*, 257–271.

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.

Coltheart, M., E. Davelaar, J. T. Jonasson, and D. Besner (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance 6*, Hillsdale, NJ. Erlbaum.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of nonnative sequences. *Journal of Phonetics 34*(1), 104–137.

Dupoux, E., K. Kakehi, Y. Hirose, C. Pallier, and J. Mehler (1999). Epenthetic vowels in japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance 25*, 1568–1578.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. Ph. D. thesis, Northwestern University.

Frisch, S., J. Pierrehumbert, and M. Broe (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory 22*, 179–228.

Goldwater, S. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory, Stockholm University*.

Greenberg, J. H. and J. J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *Word 20*, 157–177.

Hahn, U. and T. M. Bailey (2005). What makes words sound similar? *Cognition 97*, 227–267.

Hahn, U., N. Chater, and L. Richardson (2003). Similarity as transformation. *Cognition 87*, 1–32.

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.

Haunz, C. (2007). *Factors in loanword adaptation*. Ph. D. thesis, University of Edinburgh.

Hay, J., J. Pierrehumbert, and M. Beckman (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, and R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.

Hayes, B. (1999). Phonetically-driven phonology: The role of Optimality Theory and inductive grounding. In M. Darnell, E. Moravscik, M. Noonan, F. Newmeyer, and K. Wheatly (Eds.), *Functionalism and Formalism in Linguistics, Volume I: General Papers*, pp. 243–285. Amsterdam: John Benjamins.

Jurafsky, D. and J. H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. NJ: Prentice Hall.

Kruskal, J. B. (1983). An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 1–44. Reading, MA: Addison-Wesley.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. Technical report, Speech Research Laboratory, Department of Psychology, Indiana University.

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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*, 39–57.

Pertz, D. L. and T. G. Bever (1975). Sensitivity to phonological universals in children and adolescents. *Language 51*, 149–162.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of the North East Linguistics Society 23*, 367–381.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception and Psychophysics 60*, 941–951.

Pullum, G. K. and B. C. Scholz (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review 19*, 9–50.

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

Sendlmeier, W. F. (1987). Auditive judgments of word similarity. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 40*, 538–546.

Steriade, D. (1997). Phonetics in phonology: The case of laryngeal neutralization. UCLA ms.

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.

Adam Albright
MIT Linguistics and Philosophy
77 Massachusetts Ave, 32-D808
Cambridge, MA 02139
Email: albright@mit.edu