Explaining Sonority Projection Effects

Robert Daland[a], Bruce Hayes[a], Marc Garellek[a],
James White[a], Andrea Davis[b], Ingrid Norrmann[c]

[a] Department of Linguistics, UCLA; [b] Department of Linguistics, University of Arizona; [c] Department of Spanish & Portugese, UCLA

Abstract

Sonority projection refers to behavioral distinctions speakers make between unattested phonological sequences on the basis of sonority. For example, among onset clusters, the well-formedness relation [bn] > [lb] is observed in speech perception, speech production, and nonword acceptability (Albright, in preparation; Berent, Steriade, Lenertz, & Vaknin, 2007; Davidson 2006, 2007). We begin by replicating the sonority projection effects in a nonword acceptability study. Then we evaluate the extent to which sonority projection is predicted by existing computational models of phonotactics (Coleman & Pierrehumbert 1997; Hayes & Wilson 2008; *et alia*). We show that a model based only on lexical statistics can explain sonority projection in English without a pre-existing sonority sequencing principle. To do this, a model must possess (i) a featural system supporting sonority-based generalizations and (ii) a context representation including syllabification or equivalent information.

1 Introduction

The Sonority Sequencing Principle (SSP) is the cross-linguistic generalization that the most well-formed syllables are characterized by a sonority rise throughout the onset to the nucleus, and a fall from the nucleus throughout the coda (Sievers 1881; Jesperson 1904; Hooper 1976; Steriade 1982; Selkirk 1984). For example, the onset [bn] is more well-formed than the onset [lb] because the former contains a small sonority rise (obstruent to nasal) and the latter contains a large sonority fall (liquid to obstruent). A fundamental goal of phonological theory is to understand broad generalizations like the SSP.

A complete understanding involves answers to the following questions. Is the SSP synchronically active in speakers' grammars, or a diachronic byproduct of physical factors governing speech perception and production, or some combination of both? If the SSP is a part of speaker's grammars, is it innate, or learned, or some combination of both? And if learned, from what? How is knowledge of the SSP to be formally characterized? And how is it deployed during speech production and speech perception?

What is known at present is that the SSP is synchronically active in speakers' grammars (although this does not rule out diachronic factors in addition). The most recent form of evidence to support this conclusion is the existence of sonority projection effects – responses to novel stimuli that vary depending on the extent of the sonority violation. In particular, strongly SSP-violating clusters are more likely to be produced and perceived with vowel epenthesis, e.g. [lb] → [lɘb] is more likely than [bn] → [bɘn] (Davidson 2006, 2007; Berent *et al*. 2007).[1] These are termed projection effects (in the sense of Baker, 1979) because the offending clusters are systematically and *equally* absent from speaker's input, and yet speakers appear to *differentiate* some clusters as less well-formed than others.

What is not known is how the SSP comes to be a part of speakers' grammars; in fact, this is controversial in the literature, and a principal goal of this paper is to contribute to the debate. This paper focuses on the following questions:

1. What properties must any phonotactic model have in order to predict sonority projection effects?
2. How do speakers come to possess knowledge of the SSP? If it is not innate, upon what kind of experience is it based?

[1] Lisa Davidson (p.c.) and Donca Steriade (p.c.) suggest that 'sonority' may not be a true phonological primitive, but rather consists of a host of phonetic factors. For example, Davidson (2010) proposes an account of her (2006) production effects in terms of articulatory (mis)coordination. We imagine that perceptual epenthesis (Dupoux et al., 1999; Berent et al., 2007) can be accounted for similarly to the production account outlined in Wilson & Davidson (in press): the phonotactic probability of the epenthetic parse is so much higher than the intended parse that the listener simply misconstrues the token as containing the phonotactically acceptable sequence. These are all important questions that deserve further research, but for the purposes of the present paper, what is crucial is that behavioral variation can be predicted for clusters speakers have no experience with, and this variation generally lines up with traditional definitions of sonority. We believe the experimental and modeling results presented here are just as compelling whether sonority is interpreted as a true phonological primitive or as a cover term for a variety of phonetic properties.

We defer the question of model properties for the moment. As for how something like the SSP comes to be known, we distinguish *lexicalist* theories – in which the SSP is projected from the lexicon – from *universalist* theories. The lexicalist hypothesis is consistent with a body of work demonstrating other phonotactic generalizations that are projected from the lexicon (e.g. Frisch & Zawaydeh 2001; Hay, Pierrehumbert, & Beckman 2003). The universalist hypothesis comes in two forms. The most direct form is to posit that the SSP is *innate*. The innatist approach is common currency in linguistic theory, although for many specific aspects of grammar it is difficult to find a theorist who advocates an innate explanation. The other universal approach that has been proposed is that the SSP is *phonetically grounded* – learned from experience in producing and comprehending speech, and universal because "certain basic conditions governing speech perception and production are necessarily shared by all languages, experienced by all speakers and implicitly known by all" (Hayes & Steriade 2004). These possibilities are schematized in Table 1.

Hypothesis	**Projected from**
lexicalist	lexicon
innatist	Universal Grammar
phonetically grounded	speech perception/production experience

Table 1. Explanations for the Sonority Sequencing Principle.

At present, the lexicalist hypothesis is the dominant explanation for phonotactic knowledge: evidence from a variety of methodologies converges on the conclusion that the lexicon is an important seat of phonotactic generalizations. For example, the strength of gradient OCP-Place effects in nonword acceptability judgements is predictable from lexical type statistics (Frisch & Zawaydeh 2001; see also Coleman & Pierrehumbert 1997). As another example, nonword repetition accuracy is believed to index phonotactic proficiency (Coady & Evans 2008) and is strongly predicted in children by their vocabulary size, as consistent with the view that the phonotactic grammar is projected from the lexicon (Edwards, Beckman, & Munson 2004; see also Hay *et al*. 2003). The question is not whether the lexicon is a source for phonotactic generalizations, but whether it is the *sole* source.

To show that there is some other source, it would be necessary to find a particular phonotactic generalization and demonstrate that it cannot be projected from the lexicon. Just such an argument has been made for the SSP, in stronger or weaker forms, by several authors. The argument goes as follows: Lexicalist models assign well-formedness on the basis of lexical frequency. Unattested clusters have a frequency of 0. Therefore, lexicalist models should classify all unattested clusters as ungrammatical, and crucially, *equally* ungrammatical. In other words, they should fail to pick out some (strongly SSP-violating) clusters as *more* ungrammatical than other (weakly SSP-violating) ones. Sonority projection effects occur, and so lexicalist models are unable to account for them. This argument is made explicitly by Ren, Gao, & Morgan (2010, abstract):

> The sensitivity to the SSP can hardly be accounted for by lexical statistic factors because Mandarin syllables have no onset clusters and no coda consonants with the exception of [n] and [ŋ], so all the stimuli in our experiments were alien to them. The sensitivity cannot be explained by phonetic confusions either, because similar sensitivity has also been found in reading tasks (Berent 2009). The two findings shed light on … basic questions of Generative Grammar by indicating that the SSP, as a Universal Principle, may constitute a part of human linguistic knowledge.

Berent *et al*. (2007) argue similarly. They show that a particular lexical model, the Vitevitch & Luce (2004) Phonotactic Probability Calculator, has no statistically significant correlation with the results of their sonority projection study. They conclude (pp. 624-625):

> Our findings demonstrate that English speakers manifest sonority-related preferences despite the lack of lexical evidence, either direct (i.e., the existence of the relevant onsets in the English lexicon) or indirect (the statistical co-occurrence of segments in English words).

Experimental results along these lines (see also Berent, Lennertz, Jun, Moreno, & Smolensky 2008; Albright 2009) constitute intriguing evidence for the hypothesis that the SSP is not projected from the lexicon. In the theoretical taxonomy of Table 1, they may be taken as supporting either the innate or the phonetically-grounded hypotheses. However, as Berent et al. (2007, p. 624) point out, the argument relies on the failure of particular statistical models to predict the result, and there is no guarantee that other models will similarly fail. It is this point that we pursue here.

Lexicalist models assign well-formedness on the basis of lexical frequency. The key question, however, is frequency of what? *Segments* are a natural starting point for phonological analysis, and there is abundant evidence that they represent a psychologically important level of representation. However, segments are not the only representation available for analysis, and from a phonological standpoint, they are not necessarily even the best one. An alternative, noted by Berent et al., is to consider models that employ features, *i.e.* acoustic and/or articulatory properties that are shared by natural classes of segments.

If a model is limited to counting segments, then it is true that, for example, the onsets [tl] and [lt] are equally unattested. However, from a featural perspective, the onset [tl] receives more lexical support than the onset [lt]. There are many attested onset clusters that are featurally similar to [tl], e.g., [pl], [kl], [tr], [tw], [sl]. In contrast, there are no attested onset clusters that are equally similar to [lt]. A lexicalist model that generalizes across multiple featural levels of abstraction might distinguish degrees of well-formedness between these clusters on this basis, even though the segmental frequency of each cluster is 0. Indeed, at least two lexicalist models have been proposed that do generalize on the basis of features: the Hayes & Wilson (2008) Phonotactic Learner and Albright's (2009) featural bigram model. However, there is as yet no published work assessing feature-based computational models for sonority projection (though see Albright, in preparation).

Thus, the goal of this paper is to test a variety of published computational models of phonotactics on this case of sonority projection effects. The value of a direct comparison on the same stimuli is that we may gain clear insight on what model properties are responsible for success and failure on this particular phonotactic domain – which may inform our understanding as to what collection of properties the next generation of models should have.

In order to assess the predictive utility of a model, it is necessary to have human behavioral data for the model to explain. In this case, the focus is sonority projection effects, and so we begin the paper by collecting nonword acceptability ratings with nonwords whose onset clusters vary in the extent of SSP-violation. As a matter of general interest, we also included nonwords with frequently attested onsets (like [bl]) and marginally attested onsets (like [bw]).

With nonword acceptability data in hand, the paper will proceed to the modeling stage. We implement a number of computational models of phonotactics described in the literature, specifically:

- classical bigram model (Jurafsky & Martin 2009)
- featural bigram model (Albright 2009)
- syllabic parser (Coleman & Pierrehumbert 1997)
- Phonotactic Learner (Hayes & Wilson 2008)
- Phonotactic Probability Calculator (Vitevitch & Luce 2004)
- Generalized Neighborhood Model (Bailey & Hahn 2001)

The adequacy of the models is assessed by linear regression against the nonword acceptability data.

To anticipate briefly, we find that some published models exhibit considerable success in predicting sonority projection effects. The key findings are discussed in depth later; for now they may be summarized as follows: *a lexicalist model can and does predict sonority projection effects if it has (a) the capacity to represent sonority, and (b) a representation of phonological context that is rich enough to represent the expected sonority level*. In other words, lexicalist models exhibit sonority projection when they are equipped with the representations and architecture necessary to do so. This work supports a lexicalist account of the SSP.

The paper is structured as follows. In Section 2, we describe two experiments collecting nonword acceptability judgements from the Mechanical Turk, an online labor forum. In Section 3, we give brief descriptions of the computational models tested here, none containing the SSP as a bias. In Section 4, we describe the results of computational modeling; each model was trained on the same English lexicon and then assessed on its ability to predict human judgements for unattested clusters varying in their degree of SSP-violation. In Section 5, we discuss the empirical findings of this work and their theoretical implications.

2 Sonority Projection in Acceptability Ratings

In this section we describe a nonword acceptability judgement experiment with nonwords that were designed to vary in the level of SSP violation. We begin with a summary of sonority scales, followed by a brief description of the Mechanical

Turk. The nonwords are then described, followed by the acceptability experiment. The experiment had two conditions: in the first condition, participants rated forms on a Likert scale; in the second condition, participants compared two forms and selected the better choice. The section concludes with a theoretical discussion of sonority projection, and a methodological comparison of the sensitivities of Likert rating versus comparison.

2.1 Sonority scale

To determine whether participants exhibit sonority projection (and whether phonotactic models can explain it), it is necessary to have an independent measure of sonority. A number of *sonority scales* have been proposed in the literature (e.g. Steriade 1982; Selkirk 1984; Clements 1988; Parker 2002), generally[2] having the following properties:

- each segment has a sonority value represented by an integer
- segments are grouped into sonority classes sharing the same sonority value
- the minimally sonorous class has a sonority value of 0
- sonority increments by 1 between classes

The rise of a sequence XY is defined as sonority(Y)-sonority(X). Then the SSP can be formalized by defining a threshold for acceptable rises, e.g. "onsets must have a rise of at least 2" implies that [bl] is acceptable so long as *sonority(l)-sonority(b)* ≥ 2. This type of formulation has proven remarkably successful in delimiting onset inventories cross-linguistically (see references above), and is what justifies the assignment of particular integer values to particular segment classes.

Scales proposed in the literature differ chiefly in granularity. Elaborated scales such as Selkirk (1982) distinguish obstruent voicing and manner, vowel height, and rhoticity. We selected the coarse-grained scale in Clements (1988): *obstruents* (0) << *nasals* (1) << *liquids* (2) << *glides* (3) << *vowels* (4). This scale makes only uncontroversial distinctions representing the consensus of the phonological community.[3]

2.2 The Mechanical Turk

The Mechanical Turk (https://www.mturk.com) is an online labor forum provided by Amazon.com. It was used because it offers a quick and easy way to conduct word acceptability and similar studies – the total time to complete data collection was about 1 hour for each rating method, with a cost of $3/participant + 10% commission for Amazon.com, which compares favorably with 2-3 weeks and $5-$10/participant for the equivalent laboratory study. Quality is maintained in the Mechanical Turk by the approve/reject option, and the approval threshold. Researchers may reject the work of any individual worker (and refuse to pay); they may also pre-screen by selecting workers whose approval rate is above a threshold; the recommended approval threshold is 95%. As a result, workers and the website are both directly incentivized to ensure an overall high quality of work.

---

[2] Selkirk's scale starts at 0.5 for voiceless stops. The remainder of the scale has these properties.

[3] The analyses reported in §2.5 were also computed with the richly elaborated sonority scale of Selkirk (1982). The general effects were the same: attestedness, and sonority in the unattesteds.

All participants were recruited from the Mechanical Turk using the recommended 95% approval threshold. Participants gave online consent and completed a brief language background survey surveying English proficiency, dialect, and other languages spoken. Results were retained from participants reporting 'high' English proficiency (*Likert rating*: n=2; *comparison rating*: n=12) or 'native' proficiency (*Likert rating*: n=17; *comparison rating*: n=36). The research team inspected non-native results and found that they exhibited the same qualitative patterns as natives, i.e. attesteds >> marginals >> unattesteds (see next section for details). Participants reporting 'intermediate' proficiency were paid, but their results were discarded and replaced. One (native) participant was excluded from the Likert condition for rating over 80% of the items as '1'.

2.3 Stimuli

The stimuli consisted of 96 stress-initial CCVCVC nonwords, generated by concatenating a CC onset with a VCVC tail (e.g., *pr- + -eebid = preebid*). There were 48 onsets and 6 tails. Thus, each onset was paired with 2 tails, and each tail was paired with 16 onsets (48*2 = 96 = 16*6). Eighteen clusters that never occur as English onsets (unattesteds) were chosen to vary across the whole range of sonority (e.g. [tl] involves a large sonority rise whereas [rg] involves a large sonority fall). Also included were 18 clusters that occur frequently as English onsets (attesteds) and 12 clusters that occur only rarely or in loanwords (marginals, e.g. [gw] in *Gwendolyn*, [ʃl] in *schlep*). Attested and marginal clusters were included to validate the task (participants should exhibit the preference *attesteds* >> *marginals* >> *unattesteds*) and to increase ecological validity by providing at least some test items that are plausible English words. Six VCVC tails were selected to yield almost no lexical neighbors and to avoid violating any major phonotactic constraints of English.

The list of onsets and tails is shown in Table 2, with sonority values in parentheses for the unattesteds:

Attested Onsets	**Marginal Onsets**	**Unattested Onsets (*sonority*)**	**Tails**
tw, tr, sw, shr, pr, pl, kw, kr, kl, gr, gl, fr, fl, dr, br, bl, sn, sm	gw, shl, vw, shw, shn, shm, vl, bw, dw, fw, vr, thw	pw (3), zr (3), mr (2), tl (2), dn (1), km (1), fn (1), ml (1), nl (1), dg (0), pk (0), lm (-1), ln (-1), rl (-1), lt (-2), rn (-2), rd (-3), rg (-3)	-ottiff [-ɑtɪf] -eebid [-ibɪd] -ossip [-ɑsɪp] -eppid [-ɛpɪd] -eegiff [-iɡɪf] -ezzig [-ɛzɪg]

Table 2. List of onsets and tails.

The stimuli were counterbalanced in a number of ways. Each tail appeared approximately the same number of times for each sonority range, so that, e.g. *–ottiff* would not appear more often with relatively well-formed unattested onsets. The co-occurrence of tails with onset phonemes was counterbalanced; for example, *–ottiff*

would not appear more often with an onset containing /p/. Repeated segments (e.g. *dgeegiff*) would be independently dispreferred by the OCP, so these items were avoided as much as possible. Onsets and tails were combined so as to ensure that no nonword had more than one lexical neighbor (neighbor = an existing word obtainable from the nonword by inserting, deleting, or substituting one phoneme). To control for embedded words, we avoided nonwords whose $C_1C_2VC_3$ parts formed a real word with attested and marginal onsets; for unattested onsets, nonwords whose embedded C_2VC_3 parts formed a real word were distributed across the sonority range. All of the non-words were presented in English orthography in all capital letters. To ensure that the stimuli were phonologically unambiguous, they were presented to five naïve English speakers; all non-words were pronounced as intended, suggesting that the spellings are largely unambiguous.

2.4 Design and Procedure

After giving consent and filling out the language background survey, participants completed 6 practice items, and then performed the main task. All items were presented on a single page, with radio buttons for the answers.

For the Likert rating condition, participants were instructed that they would be rating potential new words of English, that they would see multiple potential words and that they should rate them based on how likely it was that the words could become new words of English in the 21st century. The practice items were *STALLOP*, *SKEPPICK*, *THRISHAL*, *SHMERNAL*, *LBOBBIB*, *SHTHOKKITH*, and were intended to expose participants to a wide range of well-formedness. Each item consisted of a single nonword, and the responses were '1' (unlikely) to '6' (likely). Each participant rated all 96 items; four different randomizations were used to control order-of-presentation effects.

For the comparison rating condition, participants were instructed to choose the nonword that seemed more like a typical English word. The practice items were *STALLOP* vs. *THMEFFLE*, *LBOBBIB* vs. *PRIFFIN*, *THRISHAL* vs. *FTEMMICK*, *SKEPPICK* vs. *MZIBBUS*, *SHMERNAL* vs. *DWIFFERT*, and *SHTHOKKITH* vs. *THPELLOP*. Each unique nonword pair was presented to exactly 1 participant, and each participant was assigned a list of 95 items, so there were 48 participants (96*95/2 pairs = 4560 pairs = 48 participants * 95 pairs/participant). Nonword position (left or right) was counterbalanced, and participant lists were constrained to not contain any nonword more than twice.

2.5 Results

All regressions were done using the *lmer* function from the *lme4* package (Bates & Sarkar 2006) in R (R Development Core Team 2006). As a check on the task, the entire data set was regressed, using the ordered factor of attestedness (*unattested << marginal << attested*) as the fixed effect. Linear regressions with rating as the dependent variable were used for the Likert condition because the response variable is scalar; onset, tail, and participant were included as random effects. Logistic regression was used for the head-to-head condition, with each trial split into two observations corresponding to each of the nonwords; the dependent variable indicated whether the nonword was chosen (note that this splitting was

necessary because unlike normal logistic regression, the two choices change from trial to trial[4]); onset, tail, and participant preference for left/right response were included as random factors. To determine whether sonority influenced listener judgements, the data sets were restricted to trials containing only unattested clusters. Sonority was used as the fixed effect, but otherwise the regressions were the same as above (linear for Likert, logistic for head-to-head, same random effects).

Attestedness was a significant predictor of well-formedness in both conditions. Marginals (nonwords containing marginal onsets) were rated significantly higher than unattesteds (*Likert*: t=-7.4, p<1e-4;[5] *head-to-head*: z=-6.4, p<1e-9) and attesteds were rated significantly higher than marginals (*Likert*: t=10.1, p<1e-4; *head-to-head*: z=6.2, p<1e-9). To visually inspect whether there is a sonority effect in the unattested clusters, Fig. 1 plots this regression's unattested cluster random intercepts against sonority. The plot shows that sonority is an excellent predictor of the variance remaining in unattestested clusters.

[4] This analysis separates a trial into two observations, one for each nonword of the pair. The statistical model assumes these observations are independent, which is false because if one word is chosen the other must not be. This coding choice reduces the power of the method, and hence can be regarded as conservative. One alternative method was specifically designed for such circumstances, and is known as "alternative-specific condition logistic regression" (McFadden, 1974) because the choice between the alternatives is conditioned on properties that are specific to each alternative, e.g. sonority of the onset cluster. However, there does not yet appear to be an implementation that allows for random effects. Another alternative would have been to model the choice between left and right, and to include both the left and right nonwords' properties as fixed or random effects; however, this ignores the real-world structure of the problem since it assigns numerically distinct coefficients for items that occur on the left versus the right. Such a model is incorrect because, for example, [bl] is the same onset whether it occurs on the left or the right. In short, the currently available statistical methods all have minor flaws. The analysis method we selected is *implemented*, *interpretable* because there is only one set of coefficients, and *conservative* because ignoring perfect anti-correlations within a pair should reduce power.

[5] Degrees-of-freedom (*df*) are unreported because *df* is ill-defined for linear mixed-effects models (Bates & Sarkar, 2006; see also Bates' comments at https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html). Accordingly, those p-values were calculated with Monte Carlo sampling using *pvals.fnc* in the *languageR* package (Baayen. Davidson, & Bates 2008).

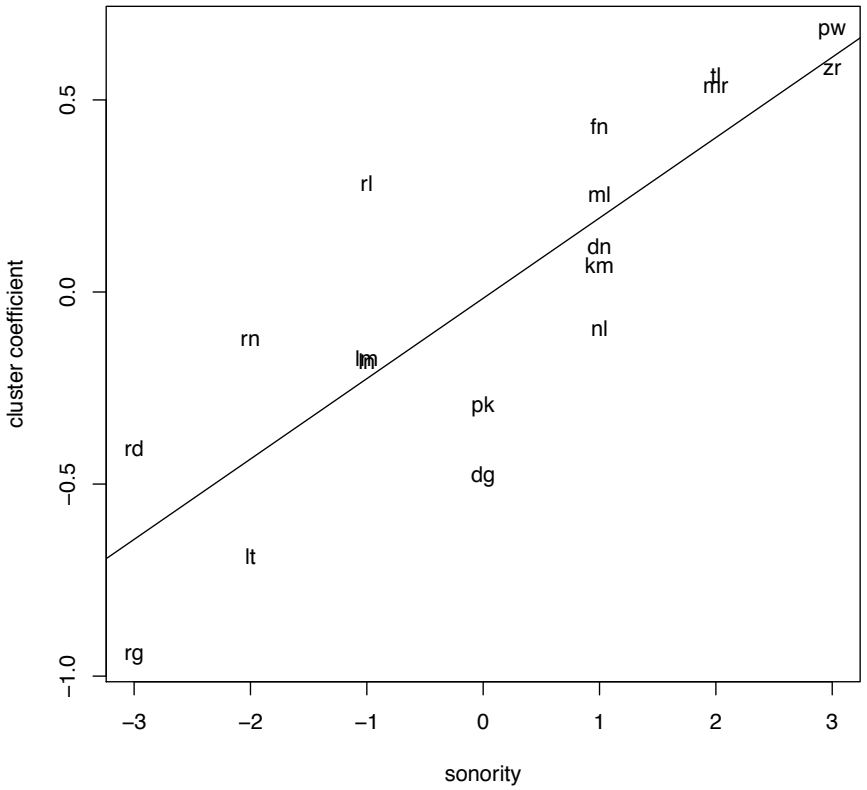


Fig. 1. Unattested cluster (random) coefficients plotted against sonority.

The sonority regression confirmed that sonority was a significant predictor of well-formedness for unattested onsets (*Likert*: t=6.2, p<1e-4; *head-to-head*: z=7.4, p<1e-12).

For modeling purposes, it will prove useful to have a canonical "acceptability score" assigned to each nonword. We define this as the proportion of comparisons trials in which a nonword was selected as better than its competitor; this value is used in preference to random intercepts from a regression model for conceptual transparency and for greater comparability to previous studies (Coleman & Pierrehumbert, 1997), though the two are highly correlated.

2.6 Discussion

The results of the nonword acceptability experiment demonstrated several important patterns. First, both the Likert rating and comparison conditions exhibited the expected effect of attestedness, with the well-formedness scale *attested >> marginal >> unattested*; this shows that participants recruited from the Mechanical Turk exhibit the same coarse behavior as laboratory participants in previous studies. Second, sonority was a significant predictor of acceptability for unattested onsets; this result is consistent with the hypothesis that speakers have internalized knowledge of the SSP, but is hard to otherwise explain. Finally, as

discussed below, while both conditions exhibited the same pattern of significant differences, the comparison condition was more sensitive for the unattested items of interest. These points are discussed in turn.

2.6.1 Inclusion of non-native speakers

The results of the present study show that at a coarse level, participants recruited via the Internet exhibit the same behavior as participants recruited through subject pools or campus flyer. Internet recruitment arguably represents a more ecologically valid sample of English speakers than a study with monolinguals, because a non-trivial percentage of speakers so recruited are early or late bilinguals. This is a potential cause for concern, as even highly proficient late bilinguals may exhibit subtle differences in judgement from native speakers (Coppieters, 1987). Note however that in this experiment the research goal is not to isolate competence of the idealized monolingual English speaker-hearer, but rather to determine whether sonority projection occurs in English nonword acceptability judgements. The other languages our participants report speaking include Dutch, French, Hindi, Mandarin, Marathi, Punjabi, and a few others; these languages are generally equally or more restrictive than English with respect to onset sonority profiles. Thus, the sonority-violating clusters in the present study are equally novel to all participants.

2.6.2 Sonority projection

The statistical modeling results showed that sonority is a significant predictor of participants' well-formedness ratings for unattested clusters. The most natural explanation for this finding is that participants have internalized knowledge of the SSP. However it is worth considering the alternative hypothesis that these results reflect some sort of orthotactic knowledge.

The orthotactic account can explain the coarse difference in rating between attested, marginal, and unattested onsets, but it fails to explain the effect of interest: sonority projection in unattesteds. The frequency of all unattested onset clusters is (by definition) 0, so they are crucially not differentiated by frequency. Moreover, the visual structure of the English alphabet does not reflect its phonology, e.g. R is more visually similar to P than to L, but more phonologically similar to L. The principled relationship between sonority and well-formedness cannot be explained by English orthotactics.

2.6.3 Sensitivity at the bottom of the scale

The pattern of significant differences was the same across the Likert rating condition and the head-to-head comparison condition. However, the comparison task was evidently more sensitive for the items of interest, the unattested onsets. One bit of evidence for this claim is that the z statistic for the head-to-head sonority comparison is greater than the t statistic for the Likert comparison, with a corresponding difference in significance (*Likert*: t=6.2, p<1e-4; *head-to-head*: z=7.4, p<1e-12). The point can be appreciated more clearly in Fig. 2, which plots raw onset averages from the comparison condition against the Likert condition. (The onset 'average' for the comparison condition is defined as the proportion of competitions won by nonwords containing the cluster.)

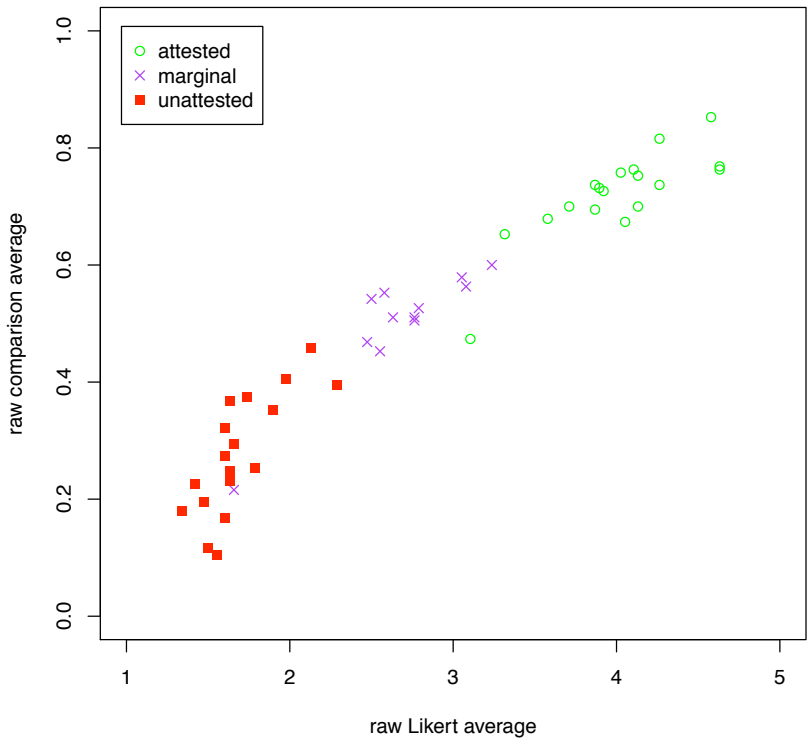


Figure 2. Scatterplot of comparison scores against Likert ratings

The comparison average differentiates the unattested onset clusters much better than the raw Likert average does. Presumably this occurs because the target, unattested items are concentrated at the bottom end of the well-formedness spectrum, yielding near-floor ratings for all of them. This fact suggests the following methodological point: *in nonword acceptability studies, head-to-head comparison is preferable to Likert rating whenever the stimuli of interest are concentrated at one end of the well-formedness scale*, owing to ceiling/floor effects in Likert ratings. Similar conclusions were reached in Coetzee's (2004) unpublished dissertation, and by Kager & Pater (under revision); and different but related points are addressed in Kawahara (ms); we mention this methodological point here in the hope of averting unnecessary replication-of-effort in the future.

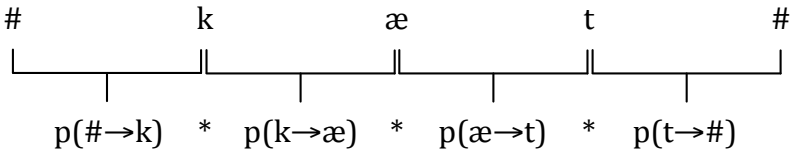
3 Computational Models of Phonotactics

Having established the human behavioral data of interest on sonority projection, we turn now to the question of explaining the judgements that humans make. In this section, we give a brief overview of six computational models that have been proposed to explain nonword well-formedness judgements. The 'training data' that will be used here is the lexicon described in §4.1, though in principle these models can train on any lexicon.

3.1 Bigram

Classical bigram models assign probabilities compositionally: the probability of the whole is the product of the probability of the sub-parts and the way they are combined. In classical bigram models, the sub-parts are bigrams, and the whole word probability is the product of the transitional probabilities. For example, 'cat' can be expressed as #kæt# (where # are boundary symbols); its bigrams are #k, kæ, æt, and t# (for detailed exposition see Jurafsky & Martin 2009; for a recent linguistic study see Goldsmith & Riggle, to appear). From these, the probability of [kæt] is calculated as in (1):

(1) *Calculating probability of* [kæt] *in a classical bigram model*



The transitional probabilities are estimated from training data using relative frequency; for example p(k→æ) is estimated by dividing the frequency of [kæ] by the frequency of [k]. This model is termed a lexical model because bigram frequencies are calculated by their type frequency in the training data – in the present study, an English lexicon.

In the natural language processing literature, where bigram and related models are heavily employed, it is considered best practice to *smooth* the transitional probabilities (Manning & Schutze 1999; Jurafsky & Martin 2009). Smoothing assigns a modest amount of probability to unseen items, so as to avoid assigning zero probability to items that happen to be absent from the training set.[6] In our implementation of the classical bigram model, we used Good-Turing smoothing (Gale & Sampson 1995).

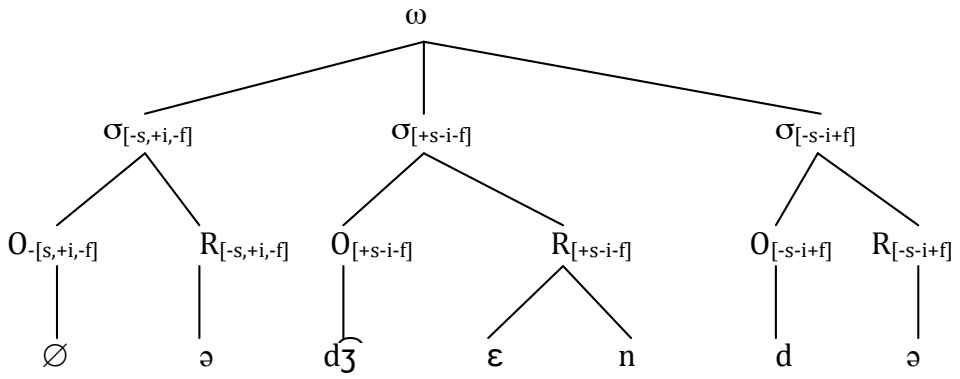
3.2 Coleman &Pierrehumbert (1997)

Coleman & Pierrehumbert's (1997) model is similar to a bigram model in that it assigns word probabilities compositionally by multiplying the probabilities of sub-parts. However, it differs in the sub-parts: words are parsed not into bigrams but into a phonologically-motivated hierarchy consisting of syllables, onsets, and rhymes. Separate counts are maintained for stressed vs. unstressed, initial vs. non-initial, and final vs. non-final syllables, for a total of eight onset and eight rhyme distributions. Just as in the bigram model, the counts are estimated from the lexicon. Our implementation differs slightly from what is described in the original paper, because our training lexicon includes more than just the binary feet on which the original implementation is based. Therefore, rather than parsing into binary feet, our implementation uses the distribution over all attested stress patterns.

[6] Smoothing is appropriate for Zipfian distributions, in which novel events continue to be observed for arbitrarily large samples. Segmental bigrams in English follow such a distribution (Daland & Pierrehumbert, 2011).

Here is an illustration of how the model computes the probability of the word *agenda*:

(2) *Calculating probability of* [əˈd͡ʒɛndə] *in the Coleman/Pierrehumbert model*



P(*agenda*) = the product of

- $P(\omega = \sigma_{[-s]}\sigma_{[+s]}\sigma_{[-s]})$ probability of a medial-stressed trisyllable
- $P(O_{[-s,+i,-f]} = \varnothing)$ probability that onset of initial stressless syllable is null
- $P(R_{[-s,+i,-f]} = [ə])$ probability that rhyme of initial stressless syllable is [ə]
- $P(O_{[+s,-i,-f]} = [d\widehat{ʒ}])$ probability that onset of medial stressed syllable is [d͡ʒ]
- $P(R_{[+s,-i,-f]} = [ɛn])$ probability that rhyme of medial syllable is [ɛn]
- $P(O_{[-s,-i,+f]} = [d])$ probability that onset of final stressless syllable is [d]
- $P(R_{[-s,-i,+f]} = ə])$ probability that rhyme of final stressless syllable is [ə]

As with the bigram model, Good-Turing smoothing was used; a separate smooth was done for each onset and rhyme distribution.

3.3 Albright (2009)

The featural bigram model is broadly similar to the classical bigram model described above. It differs in how the transition probabilities are calculated. Rather than treating each segment as a distinct, unique type, it deploys phonological features, so that each segment may be characterized by any of the natural classes to which it belongs. For example, the segment [b] can be construed as [+labial], [+consonantal], [+labial,+consonantal], [+labial,-nasal], and so on. The likelihood of a bigram is calculated from its 'best' natural-class featural description, according to (3):

(3) *Formula for selecting featural bigrams in the Albright (2009) model*[7]

$$l(xy) = \max_{A,B} p(AB) * p(x|A) * p(y|B)$$

[7] Note that this formula assigns a likelihood distribution rather than a true probability distribution, because the values do not sum to 1. This is why l(xy) is used instead of Pr(xy).

where

A and B represent natural classes to which x and y respectively belong
p(AB) is the type frequency of natural class bigram AB in the training lexicon
p(x | A) = 1/|A| (1 over the number of segments in A)
p(y | B) = 1/|B|

The overall rationale of the model is that a word containing populous natural class bigrams are likely to be particularly well-formed, especially when the segments that instantiate the natural class form a large share of that class's population. The rest of the computation works analogously to the classical bigram model.

We ran our own implementation of the model, meant to function identically to Albright's but facilitate the use of our own feature set and training data.

3.4 Hayes & Wilson (2008)

The Hayes & Wilson (2008) Phonotactic Learner is a constraint-based learning model. Constraints are stated in the phonological vocabulary made standard by Chomsky & Halle (1968) and subsequent work. For example, the constraint *#[+sonorant, -syllabic][+consonantal] militates against word-initial [lb] clusters and similar SSP-violating forms. Just as with the featural bigram model, the features allow the Hayes/Wilson model to make generalizations over segments, including generalizations based on sonority.

To assess well-formedness, the Hayes/Wilson model employs the maximum entropy variant (Della Pietra, Della Pietra, & Lafferty 1997; Goldwater & Johnson 2003) of Harmonic Grammar (Legendre, Miyata, & Smolensky 1990; Smolensky & Legendre 2006; Pater 2009; Potts, Pater, Jesney, Bhatt, & Becker 2010). Each constraint C_i has a nonnegative weight w_i. A word x is evaluated by finding its constraint violation counts $C_i(x)$, multiplying each violation count by the corresponding weight, and taking the sum. The negative of this sum is known as the *harmony* of x, and the likelihood of x is the exponential of its harmony. To ensure this is a true probability distribution, likelihood is divided by a normalization constant that guarantees it sums to 1:

(4) *Probability of word x in Hayes/Wilson model*

$$\Pr(x) = e^{\text{harmony}(x)}/Z$$
$$\text{harmony}(x) = -\sum_i w_i \cdot C_i(x)$$
$$Z = \sum_{x \in \Omega^*} \text{harmony}(x) \qquad (\Omega^* \text{ is the set of all possible words})$$

The constraints deployed in the grammar are found by a search algorithm that attempts to identify constraints that best explain the training lexicon. The algorithm privileges constraints that are brief (few feature matrices), accurate (low expected/observed violations), and general (covering large numbers of possible forms). The number of constraints that the algorithm includes in a grammar can be set by the user. We caused the model to terminate at 400 constraints, and explored the effect of constraint number by considering sub-grammars including only the

first 100, 150, 200, 250, 300, and 350 constraints. We ran the algorithm using the software posted at www.linguistics.ucla.edu/people/hayes/Phonotactics/.

3.5 Vitevitch & Luce (2004)

The Vitevitch & Luce (2004) Phonotactic Probability Calculator is widely used in psycholinguistic research. It resembles several models described already in that it assigns a score to a word by dividing it into parts and combining their probabilities. The model is similar to a bigram model in that it uses bigrams as well their simpler cousin unigrams.

The model employs a positional representation, based on *left-to-right serial position* of segments. Separate counts are maintained for each position. For example, the probability of [b] as the first segment of a word is based on what fraction of all word-initial segments are [b]; the probability of [b] as the fourth segment of a word is based on what fraction of all fourth-position segments (in words with at least four segments) are [b], and so on. In the bigram version, analogous computations are carried out on bigrams.

The model uses a weighting system evidently intended to provide a compromise between type and token frequency. It weights unigrams and bigrams by the log of their token frequencies, which are rescaled by the total log frequency weight to get unigram and bigram probabilities.

Our implementation reflects the standard practice that has evolved in experimental work making use of this model: a unigram score is calculated as the sum of the unigram probabilities, and a bigram score is calculated analogously (see http://www.people.ku.edu/~mvitevit/PhonoProbHome.html). Note that because the sub-part probabilities are not mutually exclusive, summing in this way implies that nonword scores cannot be interpreted as probabilities.

3.6 Bailey & Hahn (2001)

The Generalized Neighborhood Model (Bailey & Hahn 2001, hereafter BH2001) is an exemplar model in which the well-formedness score of an item is determined directly from the lexicon, by the sum of its similarities to existing words. This is in contrast to the other models discussed above, in which a grammar is first projected from the lexicon, and then well-formedness is evaluated by the grammar.

The similarity of a nonce word ω_i to an existing word ω_j is calculated from the *string-edit distance* d_{ij}. String-edit distance is calculated from the number of insertions, deletions, and substitutions need to change ω_i to ω_j. As in the original paper, we used an insertion and deletion cost of 0.7, and the proportion of shared natural classes (Frisch, Broe, & Pierrehumbert 1997) as the substitution cost. The similarity of ω_i to ω_j is given by $\exp(-D \cdot d_{ij})$, where D is a scaling factor. Similarly to Vitevitch & Luce (2004), log token frequency weighting was included, although this model adopts a more complicated quadratic weighting scheme. The total score for a form is calculated by summing similarities; we differ slightly from BH2001 by summing over the entire lexicon, an operation that was not computationally feasible in 2001. The full formula is given in (5):

(5) *Nonword acceptability score in Bailey and Hahn (2001)*

$$\text{score}_i = \sum_j (A\ logf_j^2 + B\ logf_j + C)\cdot\exp(-D\cdot d_{ij})$$

where

logf_j = log (token frequency of ω_j + 2)

D = 5.5;[8] d_{ij} is the string-edit distance

Owing to differences between the training lexicon and test items here versus in BH2001, we considered several sets of free parameters:

Label	A	B	C	basis
fig	-.845	3.78	-2.89	estimated from figure in BH2001
oral	-0.47	2.02	-.289	oral task in BH2001 (Bailey, pc)
writ	-.615	2.767	-1.82	written task in BH2001 (Bailey, pc)
lin	0	0	1	no frequency weighting

3.7 Summary

Each of the models discussed in the preceding section is a *learning* model. It is trained on a lexicon of a language and assigns scalar well-formedness values based on a grammar projected from the lexicon (or based on the lexicon itself). A summary of models' properties is given Table 3.

model	output	based on	from
bigram	probability	segmental bigrams	lexicon → grammar
syllabic parser	probability	syllabic constituents	lexicon → grammar
featural bigram	likelihood	featural bigrams	lexicon → grammar
Phonotactic Learner	likelihood	featural constraints	lexicon → grammar
Phonotactic Probability Calculator	scalar	positional bigrams	lexicon → grammar
Generalized Neighborhood Model	scalar	string-edit distance	lexicon

Table 3. Summary of model properties

For the models whose outputs have a probabilistic interpretation (the first four in Table 3), the outputs were log-transformed. This was partly done for comparison with well-formedness ratings since Coleman & Pierrehumbert (1997) found that nonword log-likelihoods were linearly related to human acceptability judgements; it was also simpler, as the underlying computations are actually performed in the log domain. No such log transform was applied to the scores for Vitevitch & Luce (2004) and Bailey & Hahn (2001), both for greater comparability to existing studies, and since these values are scalars that do not have a probabilistic interpretation.

[8] The approximate value of D was kindly shared by Todd Bailey, as were the A/B/C values for the *oral* and *writ* set. The *lin* setting was recommended by a reviewer.

4 Modeling nonword acceptability judgements

4.1 Training on an English lexicon

The models described in section 3 are *lexicalist learning* models, meaning that the well-formedness scores they assign are directly or indirectly projected from the lexicon. For a fair comparison, it is necessary to train all models on the same lexicon. This subsection describes the training lexicon.

Our goal was to create a representative dictionary of the words likely to be known to the participants. We used the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) transcriptions, selecting only those words that have a frequency of at least 1 in the CELEX wordform database (Baayen, Piepenbrock, & Gulikers 1995). From this set, we removed compounds, residual inflected forms, and forms created by highly transparent processes of morphological derivation, yielding a set of 18,612 words in phonemic transcription.

Two versions of the training set were used. In one, syllabification was lexically specified by annotating consonants as belonging to the coda or not.[9] In the other, coda position was not distinguished in the lexical form. The phonemes of the training set were supplemented by a feature chart. (The lexicon and features charts are available from the corresponding author upon request.) The feature chart was used by the featural bigram model, the Phonotactic Learner, and the Generalized Neighborhood Model. For these featural models, annotated coda consonants were featurally identical to their onset cousins except they were marked [+rhyme]; for the non-featural models, annotated coda consonants were counted as distinct atomic symbols, i.e. onset [b] was just as distinct from coda [b] as it was from [l].

4.2 Method

Each model was trained on the training lexicon and was then tested on the set of nonwords used in Experiments 1 and 2. Training consisted of estimating model parameters as described in §3.1-6. Testing consisted of assigning a well-formedness value to each nonword stimulus.

4.3 Results

To get a broad overview of model performance, we calculated for each model the correlation of its well-formedness score with the empirically derived well-formedness score from the experimental head-to-head data in §2.5. These correlations are shown in Table 4. The focus of this paper is on sonority projection, so what is of most interest is a model's ability to predict variation among the subset of *unattested* items. However, for completeness and general intellectual interest, we also computed correlations for the attested and marginal subsets, as well as the entire data set. These are reported in Table 4.

	syllabification		no syllabification

[9] Syllabification was assigned using the maximum onset principle (Selkirk, 1982): medial consonant sequences were parsed with the longest onset that occurs word-initially. Given that these are learning models, it is reasonable to wonder how the hidden structure of syllabification is learned. We leave this issue for future research.

model	*attested*	*marginal*	*unattested*	*overall*		*attested*	*marginal*	*unattested*	*overall*
albright	**0.21**	**0.03**	**0.55**	**0.51**		0.13	-0.07	0.18	0.26
bigram	**0.19**	**0.16**	**0.22**	**0.78**[10]		0.23	0.01	-0.14	0.50
coleman	**0.35**	**0.31**	**-0.01**	**0.55**		--	--	--	--
gnm.fig	0.07	0.25	-0.29	0.15		0.06	0.24	-0.32	0.08
gnm.oral	0.28	0.23	-0.25	0.28		0.26	0.23	-0.28	0.21
gnm.writ	0.17	0.24	-0.27	0.22		0.16	0.24	-0.30	0.15
gnm.lin	**0.32**	**0.23**	**-0.22**	**0.31**		0.30	0.22	-0.26	0.24
hw100	**0.00**	**0.02**	**0.76**	**0.83**		0.00	-0.31	0.79	0.68
hw150	0.00	0.06	0.69	0.82		0.00	0.04	0.67	0.75
hw200	-0.09	0.03	0.64	0.80		0.00	0.05	0.69	0.77
hw250	-0.09	0.13	0.64	0.84		0.00	0.00	0.70	0.80
hw300	-0.39	0.04	0.54	0.80		0.00	-0.02	0.70	0.81
hw350	-0.39	0.03	0.51	0.80		0.00	-0.10	0.67	0.81
hw400	-0.39	0.04	0.52	0.81		0.00	0.00	0.68	0.80
vl.uni	0.27	0.11	0.38	0.43		0.30	0.19	0.34	0.36
vl.bi	**0.30**	**0.06**	**0.27**	**0.56**		0.30	0.08	0.22	0.54

Table 4. Correlations of model ratings with Experiment 2 scores. Key: *albright* = featural bigram; *bigram* = classical bigram; *coleman* = Coleman & Pierrehumbert (1997); *gnm.set* = Generalized Neighborhood Model with parameter *set*; *HW[n]* = Phonotactic Learner with *n* constraints; *vl.uni/bi* = Phonotactic Probability Calculator, unigram and bigram models, respectively. "Good" model correlations are bolded (see text for details).

Models are arranged in the rows, with members of the same 'family' adjacent to one another. The columns are divided into two groups, with syllabified training/testing on the left, and unsyllabified input on the right. The columns represent the subset of the data being regressed, and the entries in each cell represent the correlation. For example, the top leftmost numerical cell indicates that the syllabified featural bigram model ratings had a 27% correlation with human judgements on the nonwords with attested onsets. These correlations provide a convenient macro-level summary of the models' predictions.

To simplify further analysis, from each family we selected a 'best' model which in our judgement represented the best or near-best performance of that family. For example, HW100 (syllabified) was selected from the Hayes/Wilson family because it had the (near-)highest attested, unattested, and overall correlations. The intention is to focus in on the most informative comparisons – those in which we can be sure relatively poor performance is not simply the result of an unfortunate choice of parameters for a model. Put another way, it is easier to understand 6 data series than 30, and since many of the data series are parametric

[10] The 'overall' score includes variation within *and across* subsets. For example, the bigram model does not do well at distinguishing unattested items from one another (low 'unattested' correlation) but it does distinguish unattesteds as a class from attesteds as a class (high 'overall' correlation).

variants, it is better to just focus on the 6 'best' ones. Fig. 3a-f plots model nonword predictions against the comparison judgements from the experiment.

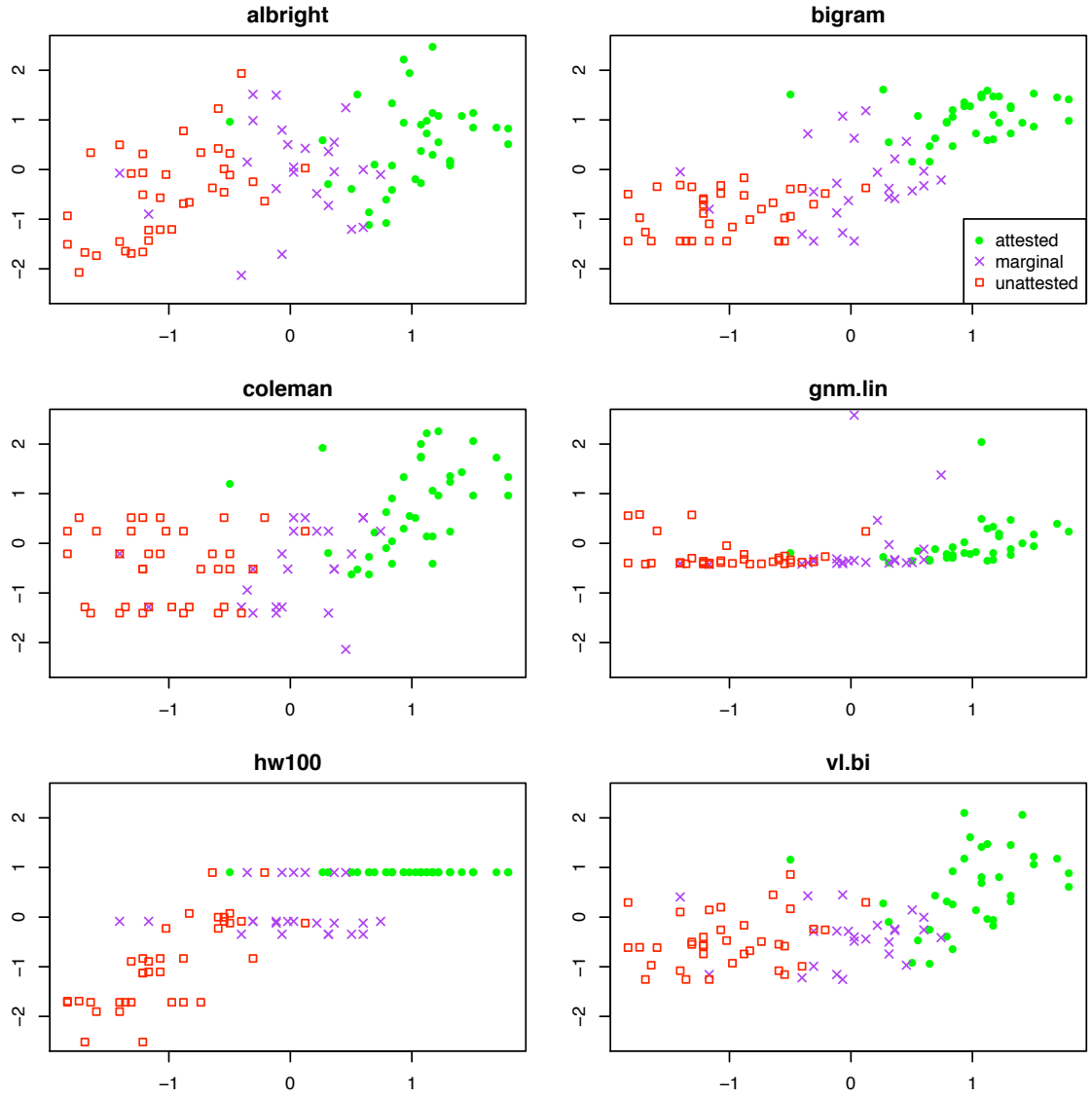


Figure 3 Model nonword predictions versus human judgements. *x*-axis: well-formedness score as determined by head-to-head comparison data (*z*-transformed); *y*-axis: model score (*z*-transformed). Each point represents a nonword. Each 'best' model is plotted in a different pane: *albright* = featural bigram model; *bigram* = classical bigram model; *coleman* = Coleman & Pierrehumbert (1997); *gnm.lin* = Generalized Neighborhood Model with no frequency weighting; *hw100* = Phonotactic Learner; *vl.bi* = Phonotactic Probability Calculator

4.4 Discussion

Several points emerged from the results of the modeling study. The most significant finding for linguistic theory as a whole was that there exist lexical models that explain sonority projection, i.e. predict sonority-related variation in human behavior for unattested phonological sequences. Among the models tested here, the models which were most effective at modeling sonority projection effects were the Phonotactic Learner (Hayes & Wilson, 2008) and Albright's (2009) featural bigram

model; we will argue that sonority projection owes to a featural representation of sonority and a rich enough representation of context to track the expected sonority profile. Another finding of interest was that no current model excelled across the well-formedness spectrum, i.e. the models that were best on unattested onsets were not best on attested onsets. These points are discussed in detail below.

4.4.1 Sonority projection is possible from the lexicon alone

The most theoretically significant finding of the present study is that sonority projection is achieved by a number of published lexicalist phonotactic models. This finding directly contravenes previous claims in the literature, such as the following passage from Berent *et al*. (2007, pp. 624-625):

> Our findings demonstrate that English speakers manifest sonority-related preferences despite the lack of lexical evidence, either direct (i.e., the existence of the relevant onsets in the English lexicon) or indirect (the statistical co-occurrence of segments in English words).

Berent and colleagues found sonority projection effects in perception, which is a substantial contribution to the field, because it unequivocally demonstrates that the SSP is a part of speakers' synchronic knowledge. What we disagree with is the claim that there is no lexical evidence for the sonority-based preferences. It is true that, for example, *lb* and *tl* are equally unattested as English onsets, but there are many onset clusters that are featurally similar to *tl* whereas there are none that are so featurally similar to *lb*. A lexicalist model that is equipped to make generalizations on the basis of features should in principle be able to explain sonority projection, and we have shown here that this is exactly what happens.

From the belief that there is no lexical support for sonority projection, Berent and colleagues draw the inference that listeners must possess some form of universal knowledge, whether it arises from "inherent preferences of the language system" (p. 593) or knowledge that is "induced from phonetic experience" (p. 625). Our results show that the inference of universal, non-lexical phonotactic knowledge does not follow as a logical necessity – although it may still be correct. In short, the ability of a lexical model to explain sonority projection effect bears on foundational issues of our field, because it refutes a powerful argument for the existence of universal phonotactic knowledge.

4.4.2 Model properties needed for sonority projection

Beyond the sheer fact that sonority projection occurs, it is of interest to know why some models exhibit it and others do not. We will argue that what is needed is the ability to capitalize on two representational properties: a sufficiently rich representation of phonological context (e.g. syllabification), and a sufficiently rich representation of sonority itself (e.g. features). The syllabified featural bigram model exhibits sonority projection. However the unsyllabified featural bigram models does not, so removing syllabification inhibits sonority projection. Similarly, the syllabified classical bigram model does not, so removing featural generalization also inhibits sonority projection. Thus, sonority projection requires both properties.

Phonological context. To express sonority restrictions, a model needs to be able to distinguish contexts that constrain the sonority profile, e.g. it should be rising word-initially. Models trained on syllabified data can do this, since they are told whether a consonant sequence is parsed as an onset, a rhyme, or as a heterosyllabic cluster. With this information, such models are in a position to inductively track the sonority profiles characteristic of these three contexts, and characterize well-formedness of these configurations when they are filled by particular segments.

We illustrate using the specific example of the featural bigram model. A cluster like [lt] is perfectly acceptable in English when it is not an onset cluster, e.g. *halt*, *Elton*. When syllabification is made available to the model, it should be able to distinguish the unacceptable onset cluster from the acceptable coda and heterosyllabic clusters. Indeed, when it is trained with syllabified data, the Albright model achieves a correlation of r=.55 with human judgements for unattested onsets. The correlation drops to r=.18 when the same model is trained on unsyllabified data (we will show later that this level of correlation arises merely from modeling tails). Since the only difference between these two cases is the presence of syllabification, it follows that the contextual information represented by syllabification caused the difference. In other words, syllabification provides a sufficiently rich representation of the context as to allow Albright's model to represent the expected sonority contour.

It is possible for a model to succeed without an explicit representation of syllabification. In particular, with a sufficiently large number of constraints the Phonotactic Learner achieves roughly equivalent performances on syllabified or unsyllabified data. We believe this owes to the fact that the Phonotactic Learner allows trigram constraints. English trigrams provide a level of phonological context that is more specific than syllabification; for example a trigram model can use structural descriptions of the form [$x\ y\ C$] and [$x\ y\ \#$] in place of $x_{coda}y_{coda}$, as was done in Chomsky and Halle (1968). At the same time, scaling up a model to trigrams has its own costs in terms of sparseness of data and computational complexity (Jurafsky & Martin 2009).

In summary, what a model needs is some representation of phonological context that is sufficiently rich as to track the expected sonority contour (see Kager & Pater, under revision, for another study concluding that phonotactic models must represent syllabification). Explicit syllabification is an especially simple and effective means of doing this, as evident from the fact that nearly every model does better on nearly every subset of the data when it has access to syllabification.

Phonological features. In addition to phonological context, a model needs a system of phonological features. The rationale for this claim is very simple: in order to make generalizations on the basis of sonority, a model must be able to make generalizations, and it must have an explicit representation of segments' sonority. Phonological features perform both of these functions. Features represent inherent generalizations, because the presence or absence of a feature represents an underlying acoustic or articulatory property shared by a natural class of segments.

And many of the features commonly used in generative phonology pertain to sonority – for example [+sonorant] segments like [n], [l], and [w] have a relatively open vocal tract providing support for formant resonance, rendering them more sonorous than obstruents like [s], [t], and [ʧ]. Phonological features are a theoretically convenient way to provide for sonority-based generalizations, because they organize segments into classes on the basis of sonority, and are independently motivated.

The necessity of an explicit representation of sonority can be illustrated by a comparison between the classic bigram and the featural bigram models. These two models differ principally in whether they are designed to generalize on the basis of featural similarity, or to stick narrowly to segmental biphone probabilities. The two models achieve comparable performance on the attested clusters (classic: r = .19; featural: r = .21), but differ substantially on unattested clusters (classic: r = .22; featural: r = .55). Since the primary difference between these models is whether they make featural generalizations, and the featural bigram model outperforms the classical bigram model on the unattesteds, it is evidently the ability to make feature-based generalizations that *causes* the difference. In other words, a featural representation of sonority and the capacity to make feature-based generalizations are responsible for sonority projection in Albright's (2009) model.

Extant models that lack a feature system, such as the classic bigram model and the syllable parser, treat segments or other prosodic constituents as atomic units. These models do not 'know' that [z] is less sonorous than [l], and so to them there is no or little principled distinction between, e.g., the onset clusters [zl] and [lz]. This seems, almost as a point of logic, fatal to the enterprise of predicting sonority projection to novel clusters. And indeed, all such models achieve at best low correlations with the unattested clusters, which we will argue below arise from modeling the tails. In contrast, the models which employ a feature system – the Phonotactic Learner and the featural bigram model – are exactly the ones with the best success at predicting sonority projection in the behavioral data. Some explicit representation of sonority, such as is generated by a set of phonological features, is a crucial ingredient for making generalizations on the basis of sonority.

4.4.3 Why lexical analogy is insufficient

It is worth asking why the syllabified GNM (Bailey & Hahn 2001) does not succeed at sonority projection. After all, this model is equipped with a featural representation. As we will show in this subsection, the reason the GNM does not exhibit sonority projection is because even when syllabification is available, the GNM fails to leverage it.

The point can be illustrated most clearly with a slight idealization. We define GNM' as identical to the GNM, except that it considers only the *closest* word in assigning a score. This idealization is relevant for the nonwords in the present study because the exponent D is quite high, which means that the score is effectively controlled by whatever word(s) have the minimal string-edit distance. Because the nonwords in the present study are in sparse lexical neighborhoods, it is safe to assume that the closest word is unique. Thus for a nonword ν with the existing word ω as a neighbor, the assigned score is $GNM'(\nu) = \exp(-5.5 \bullet d_{\nu\omega})$, where $d_{\nu\omega}$ is the

string-edit distance between ν and ω. Crucially, a nonword's score is determined purely by string-edit distance to its nearest neighbor.

Now let us consider the nonword *guzu*, for which we will suppose *guru* is the closest neighbor. It is evident that the best string alignment between *guzu* and *guru* is the one in which *z* maps to *r* and all other segments match. Thus, the string-edit distance between *guzu* and *guru* is simply the substitution cost of *z* → *r*. For concreteness, let us suppose this is 0.7, which was the insertion/deletion cost in Bailey & Hahn (2001). Then the score assigned to *guzu* is GNM'(*guzu*) = exp(-5.5*.7), because 0.7 is the cost of substituting *z* → *r*.

Now let us consider the nonword *bzoker*, for which we will suppose *broker* to be the only neighbor. It is evident that the best string alignment between *bzoker* and *broker* is the one in which *z* maps to *r* and all other segments match. The string-edit distance again consists simply of the substitution cost *z* → *r*. Then the score of *bzoker* is also determined entirely by the cost of substituting *z* → *r*. It follows that GNM' will assign the same score to *bzoker* as to *guzu*, namely exp(-5.5*.7), and moreover it is clear *why* GNM' will assign the same score to both nonwords – because both differ in exactly the same *z* → *r* way from an existing neighbor.

Now let us consider these facts from a phonological perspective. GNM' assigns the same well-formedness score to *guzu* as to *bzoker*; however, *guzu* is a perfectly legitimate nonword of English, whereas *bzoker* contains an unattested, sonority-violating onset cluster. The difference in well-formedness is evidently contextual: *z* is acceptable intervocalically, but not in the onset cluster **bz*. Indeed, Albright (2009) noted that the GNM was vulnerable to items like *bzeakfast*, which overlap strongly with existing words but contain a contextually ungrammatical substitution. We have illustrated here that it is a property of the string-edit metric that the distance between *z* and *r* does not depend on context; it treats both *z*'s equally. In other words, context-insensitivity in the underlying string-edit metric implies that the GNM' is insensitive to phonological context.

For expository purposes, the insensitivity of the GNM to phonological context was demonstrated with an idealized model in which well-formedness is determined by string-edit distance to only the nearest neighbor. However, the idealized model is a good approximation to the true GNM for the nonwords in this study, and the insensitivity of the string-edit distance to phonological context holds equally true for the real, non-idealized GNM. Thus, the GNM fails to leverage the phonological contextual information that conditions sonority projection, even when it is provided in the training data.

The reader may wonder why the GNM is actually anti-correlated with human judgements on the unattesteds, rather than simply uncorrelated. The answer lies in what we refer to as *delete-initial neighbors*. A word *y* is a delete-initial neighbor to nonword *x* if *y* is the closest[11] nonword to *x*, and the initial segment of *y* is deleted in the string alignment to *x*. For example, the nonword *rteppid* has *tepid* as a delete-initial neighbor. In the GNM, the effect of a delete-initial neighbor is to give a boost to a nonword, irrespective of whether its onset is attested or unattested. It is

[11] This does not imply that the only difference is in the initial segment. There may be other changes, as long as there is no other word that is closer.

evident from inspection of the GNM pane of Fig. 3 that several nonwords with unattested onsets have such delete-initial neighbors. The relevant nonwords are *rgeebid*, *rgeppid*, *lmeebid*, *rlezzig*, *dgeppid*. These words are concentrated at the bottom end of the well-formedness spectrum, and so the GNM assigns a higher score to a few particular words that are phonologically the least well-formed. This explains why there is actually a negative correlation with well-formedness ratings: our nonword set happened to contain a number of *bzeakfast*-type items that were concentrated at the bottom of the well-formedness spectrum.

As a side note, the failure of the GNM rules out an important alternative interpretation of the experimental results. Various scholars have raised the issue that nonword acceptability judgements do not reflect pure phonological intuitions, since lexical analogy is known to play a role (Bailey & Hahn, 2001; Goldrick, in press). We have explained in detail why lexical analogy – at least as it is implemented in the GNM – fails to predict sonority projection, and actually results in anticorrelation with human judgements. This strongly suggests that lexical similarity cannot explain the sonority projection effect we observe for unattested onset acceptability.

4.4.4 The contribution of tails

A reviewer raised the concern that for non-featural models, even small correlations on the unattested items are unpredicted under our account (because they do not have featural generalization). Thus, the modest success of the Phonotactic Probability Calculator (r=.27) and the bigram model (r=.22) on syllabified training data are of some concern, as are the same models' correlations on unsyllabified training data (r=.22 and r=.18, respectively). We will show that these modest correlations arise entirely from modeling the contribution of tails, beginning with Fig. 4.

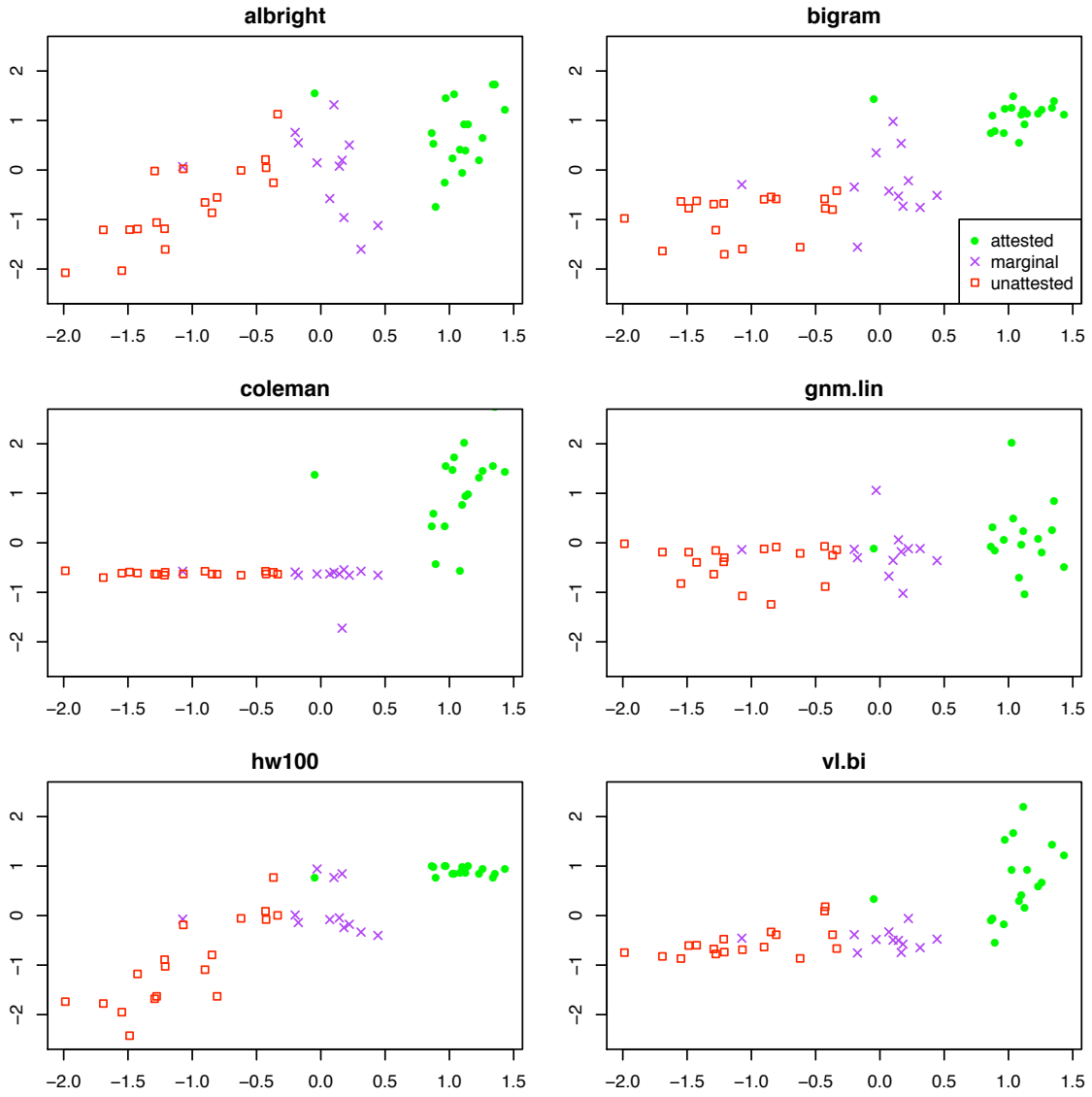


Fig. 4: Model cluster prediction versus human judgements. Each point represents a cluster. All other graph properties are as in Fig. 3.

Fig. 4 isolates the contribution of clusters in the following way. For each model, the contribution of a tail was defined as the average score assigned by the model across all nonwords possessing that tail. For each nonword, this contribution was subtracted out from the nonword score. Once the tail contributions were subtracted out, cluster scores were obtained by simple averaging. As with Fig. 3, both *x*- and *y*- coordinates were *z*-transformed for visual comparison.

It is evident in Fig. 4 that the non-featural bigram models assign essentially flat scores to all the unattested items. The correlation of a constant (model score on unattested onsets) with a variable (human judgements of the same onsets) is zero. Since the models achieve zero correlation on the onsets alone, but a modest positive correlation on the onsets plus the tails, the modest correlations must be caused by modeling the tails' contributions to well-formedness.

Indeed, not only do the non-featural model scores reflect variability from the tails, but they appear to weight the tail contribution too heavily. This conclusion follows from an important difference between human judgements and model scores. Human judgements are dominated by the onset differences. This is evident from the fact that the three attestedness categories are cleanly separated on the *x*-axis of Fig.'s 3 and 4; it is also evident from Fig. 1, which suggests that onset sonority profile is the most important predictor of nonword well-formedness. In contrast, the non-featural bigram models sometimes assign higher scores to illegal nonwords than to legal ones, as evident from the fact that the attestedness categories are *not* cleanly separated on the *y*-axes of Fig. 3. This fault cannot originate with the onset component of the model scores, since the unattested onset components are essentially flat (as shown in Fig. 4). Therefore the issue must lie with the tails. In other words, a nonword with an illegal onset and a very likely tail (e.g. *nlezzig*) can be scored better than a nonword with a legal onset and a somewhat likely tail (e.g. *sneegiff*). For example, the Phonotactic Probability Calculator scores *nlezzig* as .0040 and *sneegiff* as .0028. In non-featural models, the tail component sometimes trumps the onset component, but this does not occurs in our human judgement data.

In summary, the modest positive correlation of the non-featural models on the unattesteds can be attributed to the contribution of the tails, as evident from factoring out tail contributions (in Fig. 4). Moreover, the comparison between Fig. 3 and Fig. 4 draws out an important weakness of these models – they are *too* sensitive to the tails: the tail component sometimes trumps the onset component in non-featural model scores, but not in human judgements.

4.4.5 Predicting judgements on attested items

Although the focus of this study is the models' predictions for the unattested items, their predictions for attested onsets are of general theoretical interest as well. As shown in Table 3, the models that achieve the best performance on nonwords with attested onsets are the syllabic parser (*r*=.35), the GNM (r=.32), and the Phonotactic Probability Calculator (*r*=.30). This finding is of special interest since these are among the worst models at predicting judgements on the unattested items. In other words, the properties that are necessary for predicting judgements on unattested items are not the same as the properties necessary for predicting judgements on attested items (although there may be some overlap). Then, what is responsible for these relative successes on the attested items?

We believe the property is inherent in the design of these models: conformance to lexical type statistics (e.g. type frequency). This is the essential property that the syllabic parser and Phonotactic Probability Calculator share; and it is directly encoded in the lexical similarity measure of the GNM. For additional theoretical and empirical evidence and argumentation supporting the role of lexical type statistics, see Hay *et al*. (2003), Edwards *et al*. (2004), *inter alia*.

4.5 Summary

The models that predict sonority projection on the unattested onsets were the syllabified featural bigram model (Albright 2009) and the syllabified Phonotactic Learner (Hayes & Wilson, 2008). We argue that the model properties

that underlie this success are (i) a representation of context sufficiently rich as to distinguish expected sonority contour, and (ii) a featural representation enabling sonority-based generalizations. Models that lack either of these properties – or the ability to exploit them in the proper way – will fail at sonority projection. For example, the GNM fails because the string-edit metric is not sensitive to phonological context. The modest positive correlations in Table 6 of some models not possessing both properties can be attributed to modeling variation in the tails; comparison between Fig. 3 and Fig. 4 shows that these models overweight the contribution of tails relative to human judgements. Finally, the models which do best at predicting judgements on unattested onsets are among the worst at predicting judgements on attested onsets; apparently what is needed for attested items is conformance to lexical type statistics or lexical analogy.

5 Pushing the lexicalist account to its limits

It has been argued that sonority projection effects in Korean (Berent *et al.* 2008) and Mandarin (Ren *et al*. 2010) provide evidence against the lexicalist account. The argument runs as follows. (i) Korean and Mandarin lack onset clusters. (ii) Lexicalist accounts predict that a language must have consonant clusters in order to induce the SSP. Therefore, the lexicalist account predicts no sonority projection effects in these languages. (iii) Sonority projection effects are evident in Korean (Berent *et al*. 2008) and Mandarin (Ren *et al*. 2010). (iv) So the lexicalist account makes an incorrect prediction. We do not dispute the existence of sonority projection effects in Mandarin and Korean, or that conclusion (vi) follows if (i)-(iii) are true. However, we will show here that (i) is questionable and (ii) is false. Both languages could be analyzed as having surface obstruent-glide clusters, so it is not clear these languages are the proper test case. However, even if they are, modeling results show that sonority projection can be explained even from exposure to a CV language.

We will begin with the phonological analysis of Korean and Mandarin. We take it as uncontroversial that Korean and Mandarin allow syllables whose onset and nucleus jointly contain 3 segments, such as the family names *Choi* (Korean) and *Huang* (Mandarin). These items are traditionally analyzed (e.g. Hockett 1947, p. 223) as containing diphthongs in which the second segment is affiliated to the nucleus: [tɕuæ], [huaŋ]; under this analysis, it is true that both languages lack onset clusters. However, there are good reasons to analyze these items as having complex, obstruent-glide onsets: [tɕwæ], [hwaŋ] (Korean: Lee 1994; Mandarin: Duanmu 2000, p. 86). For example, Korean generally allows labial and coronal approximants in the onset position, but specifically disallows them before the structurally ambiguous segment in question (*[juæ], *[wiæ]); this absence has every appearance of a sonority effect, and could be taken as evidence for the complex onset analysis if the SSP is construed as regulating the onset profile specifically. Since it is *a priori* reasonable for learners to entertain the hypothesis that such sequences are surface clusters, and some aspects of the data arguably favor this hypothesis, the claim that Korean and Mandarin lack complex onsets is not really clear-cut.

Even if sonority projection were demonstrated in a strict CV language, Hayes (in press) showed this does not demonstrate the need for universal (non-lexical)

phonotactic knowledge. This point was demonstrated using simulations with the Phonotactic Learner on artificial languages called Ba and Bwa. Ba consisted of every possible CV syllable; Bwa included all CV syllables as well as all possible syllables with a stop-glide onset (hence the name Bwa). The segmental inventory, arranged by sonority class, was [*ptkbdg*] (stop) $<<_1$ [*fvsz*] (fricative) $<<_2$ [*mn*] (nasal) $<<_3$ [*rl*] (liquid) $<<_4$ [*wj*] (glide) $<<_5$ [*a*] (vowel). The integers here are for expository convenience and represent the divisions between classes on the sonority scale, as follows: [-son α] represents all segments below breakpoint α, e.g. [-son 2] represents the stops and fricatives. The Learner was endowed with two families of 'sonority-regulating' constraints, i.e. those in which some minimal difference between the initial and the final consonant of the cluster is enforced:

SSP: *[+son α][-son β]
anti-SSP: *[-son β][+son α]

Crucially, the constraints included both SSP-enforcing constraints and their exact opposites. For example, *[+son 4][-son 1] bans glide-stop clusters like *rd*; *[-son 1][+son 4] bans stop-glide clusters like *dr*. Sonority projection was identified as the presence of a gradient of well-formedness across C_1C_2 clusters, in which well-formedness increased with the sonority of C_2, and decreased with the sonority of C_1. Hayes (in press) found sonority projection in Bwa, indicating that the presence of just obstruent-glide clusters is enough to trigger sonority projection. Crucially, Hayes also found sonority projection in Ba, indicating that sonority projection from the lexicon may occur even without clusters in the input. This arose because featural generalization of the sharp sonority rise from C to V results in a sonority-dependent gradient of well-formedness for C-to-C.

In summary, the claim that Korean and Mandarin lack complex onsets is problematic. Even if this analysis is accepted as the one that learners definitely make, the existence of sonority projection in these languages does not clearly refute the adequacy of a lexicalist model. Hayes (in press) demonstrated that the Phonotactic Learner exhibits sonority projection on CV languages when it is equipped with constraints that include both the SSP and its exact opposite. The existence of sonority projection in these languages therefore does not demonstrate the need for, and existence of, universal (non-lexical) phonotactic knowledge.

6 Discussion and Conclusions

We conclude the paper with a summary of the major findings and brief discussion. In summary, we showed that

(1) Sonority projection is evident in nonword acceptability studies. Head-to-head comparison is appropriate because the target stimuli are concentrated at the bottom end of the well-formedness scale.
(2) To explain sonority projection, *any* phonotactic model must be equipped with a featural representation of sonority and a representation of phonological context such as syllabification.

(3) When so equipped, lexicalist models with the capacity for context-sensitive featural generalization can and do explain sonority projection. English well-formedness judgements on SSP-violating clusters were well-modeled by the Phonotactic Learner (Hayes & Wilson 2008) and Albright's (2009) featural bigram model.

(4) The existence of sonority projection effects in Mandarin and Korean (Ren *et al*. 2010; Berent *et al*. 2008) do not falsify the lexicalist account, as sonority projection can be explained by a lexicalist model exposed only to CV syllables.

The latter two points refute arguments against the lexicalist account that have been interpreted as powerful support for universal (non-lexical) knowledge of the SSP. Thus, the results in this paper bear on foundational issues of our field. In the remaining sections, we enlarge on some of these points.

6.1 Sonority and context are needed for sonority projection

In this paper, we have argued that a representation of sonority and context are needed for a lexicalist model to exhibit sonority projection. However, there is nothing in the theoretical arguments we made that is specific to lexicalist models. Rather, this point should obtain for all phonotactic models that predict sonority projection. To exhibit gradient sensitivity to degree of sonority violation, a model must have a gradient representation of sonority, such as a standard featural scale. And since the sonority profile intrinsically depends on at least two segments and their relation to the nearest sonority peaks, a model must represent this context in order to predict the expected sonority contour. These points do not depend on where knowledge of the SSP comes from (lexicon, phonetic experience, innate, etc..); rather they refer to properties of the representation that a model must have to adequately explain human performance.

6.2 Lexical models that generalize succeed with sonority and context

In the preceding section we summarized arguments to the effect that any phonotactic model must have a gradient representation of sonority and an adequate representation of phonological context in order to explain sonority projection effects. The most significant empirical contribution of this paper is to show that when a model has these properties, and the capacity to make generalizations on the basis of them, it may succeed at sonority projection. Our results refute the view that lexicalist models are unable to explain sonority projection and therefore humans must possess some universal knowledge of the SSP. For example, the model that Berent *et al.* (2007) used was the Phonotactic Probability Calculator (Vitevitch & Luce, 2004). We have shown here that the problem was not that its predictions were derived from lexical frequencies, but rather that it lacked the capacity for generalization based on sonority, because its computations were based on atomic representations of segments that excluded sonority.

In the present paper, we showed that models that possess both a featural representation of sonority and the capacity for feature-based generalization do in fact predict sonority projection. The relevance of these properties was empirically

demonstrated by "minimal model pairs". Albright's (2009) featural bigram model with syllabified input succeeded at sonority projection (= adequately modeled sonority-based variation in human judgements on nonwords with unattested onset clusters). In contrast, the featural bigram model without syllabification lacked an adequate representation of context, and it failed. The classic bigram lacked a featural representation of sonority from which to generalize, and it failed. Since these models are otherwise identical in their essentials to the featural bigram model with syllabification, it must have been syllabification and featural generalization specifically that caused success in the one case, and their absence that caused failure in the other cases.

In summary, we have offered both theoretical and empirical arguments for our core position: when a lexicalist model is equipped with the necessary properties of a featural representation of sonority and an adequate representation of context, it not only can but does explain sonority projection.

6.3 Synchrony, diachrony, and the SSP

We began this paper by posing the questions of what properties any phonotactic model must have, and where knowledge of the Sonority Sequencing Principle comes from. In the present paper, we have argued that synchronic knowledge of the SSP may derive from the lexicon; moreover we have shown that the English lexicon provides a great deal of support for the SSP. However, there is a significant challenge remaining under the hypothesis that the SSP does in fact derive from the lexicon: why is the SSP an apparent typological universal?

Under both the innatist and phonetically-grounded accounts, the universality of the SSP is transparently accounted for. Under the innatist account, universality is accounted for by the assumption that the SSP is part of our common human endownment of UG. Under the phonetic grounding account, sonority projection derives from the speaker-hearer's implicit knowledge of articulatory and perceptual relations, e.g. the onset *rt* is dispreferred both because of the articulatory difficulty of initiating, pausing, and then re-initiating voicing, and because the perceptual cues to the presence of a word-initial *r* are obfuscated by the following obstruent (for a summary of perceptual cueing see Wright 2004). This knowledge is universal because we all share the same articulatory and perceptual mechanism. In either case, it is clear why the SSP is universal.

In contrast, the lexicalist hypothesis is that the SSP emerges from the lexicon. Thus, the SSP should only be universal if there is some process that universally causes lexicons to prefer words that are in conformance with the SSP. The lexicalist account does not in of itself explain why lexicons are cross-linguistically structured so as to support the SSP, as the null hypothesis is that lexicons are subject to arbitrary variation (for discussion see Prince & Smolensky, 1993; Joseph, 1995), including conformance to the SSP. Since they are not, the lexicalist hypothesis must be supplemented with some diachronic hypothesis that explains why SSP-conforming words come to predominate in the lexicons of the world's languages.

The obvious candidate is Evolutionary Phonology (Blevins, 2004). EP proposes that sound patterns are "phonologized" when a gradient phonetic phenomenon is analyzed as an automatic, categorical process. A crucial aspect of EP

is that it posits that phonology is not "teleological", i.e. phonetically unnatural patterns (such as the English *k-s* alternation embodied in *electric~electricity*; Pierrehumbert 2006) can be phonologized just as easily as phonetically natural ones like nasal place assimilation (for additional examples see Baroni 2001, Koo & Cole 2006, and Kawahara 2008). Under this account, the reason that phonetically natural patterns are more likely to be phonologized is because they occur more frequently. Thus, EP differs crucially from the phonetic grounding account in locating some language universals in the *conditions* governing speech production and perception in the world, rather than the *grammar* in the minds of speaker-hearers.

A detailed proposal of how the EP could account for the predominance of SSP-conforming words in a lexicon is beyond the scope of the present paper. However, the general idea seems clear. Suppose that a lexicon begins with a number of SSP-violating words, as well as a number of SSP-conformers. As documented by Blevins (2004) and others, the phonetic factors governing speech perception and speech production might cause misperception and misproduction more frequently in the SSP-violators. For example, sonority-violating clusters might be more likely to be both produced and perceived with an epenthetic vowel even though the offending clusters are legal in the language (Berent *et al*. 2007). Such words might gradually acquire an underlying vowel that repairs the difficult cluster. Thus the lexical support for these clusters is gradually eroded, with the end result that they become unattested.

This sketch is different from the phonetically-grounded account, but not incompatible with it. Under the phonetically-grounded or innatist accounts, knowledge of perceptual and articulatory difficulty is encoded in the grammar, i.e. ill-formedness is mentally represented. Under the EP-style account, evolutionary selection favors SSP-conforming words even if ill-formedness is *not* mentally represented. This is necessary to avoid circularity – we could hardly claim that the lexicalist account avoided the need to posit universal phonotactic knowledge if it appealed to an evolutionary account that assumed universal phonotactic knowledge. Thus it is necessary that the EP account could explain the lexical universality of the SSP even without universal (non-lexical) knowledge, i.e. that SSP-conformers are evolutionarily selected for even without an intrinsic grammatical preference. However, the fact that the EP account can explain the lexical universality of the SSP without grammatical universals does not thereby imply that we believe universal accounts are incorrect. Indeed, we suspect that ultimately the SSP will prove to derive from a combination of lexical and universal (non-lexical) knowledge.

6.4 The next generation

Beyond the theoretical contribution to our understanding of the SSP, the results of the simulations here suggest a promising outlook for the next generation of phonotactic models. Recall the finding that the models that did best on the unattested onsets were not those that did best on attested onsets. Moreover, we offered informed opinions as to which properties were responsible for success on each domain. The models with featural generalization did best on unseen, ungrammatical items, while the remaining models whose computations were based

on lexical type statistics did best on nonwords whose parts were all attested. These properties are not inherently mutually exclusive.

Therefore, it is tempting to believe that a better model could be built by incorporating the best properties for both domains. For example, the Phonotactic Learner currently assigns essentially flat scores to all nonwords that are relatively well-formed. It is quite likely that this outcome derives from search biases built into the model that were explicitly intended to help it find exceptionless constraints (Hayes & Wilson, 2008), and that the model would also distinguish gradient well-formedness predictions on attested items if these biases were relaxed. Similarly, the featural bigram learner here does less well than the classical bigram on nonwords whose subparts are all attested. Presumably this occurs because the model allocates too much probability mass to featural generalization, when on attested items it would do better to simply abide by the existing lexical statistics. Analogously, the GNM could be improved by using a similarity metric that is sensitive to phonological context. In short, the model properties needed for unattested items are not mutually exclusive with the properties needed for attested items; there is nothing principled standing in the way of the next generation of models incorporating both types of properties.

References

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1), 9-41.

Albright, Adam (in preparation). Natural classes are not enough: Biased generalization in novel onset clusters.

Baayen, R. Haarald, Douglas J. Davidson, and Douglas M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.

Baayen, R. Haarald, Richard Piepenbrock, and Leon Gulikers (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533-581.

Bailey, Todd M & Ulrike Hahn (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44, 568-591.

Baroni, Marco (2001). How do languages get crazy constraints? Phonetically-based phonology and the evolution of the Galeata Romagnolo vowel system. In Adam Albright and Taehong Cho (eds.) *UCLA Working Papers in Linguistics 7*, 152-178. Los Angeles: University of California, Los Angeles Department of Linguistics.

Bates, Douglas M. & Deepayan Sarkar (2006). *The lme4 package*. http://cran.r-project.org/src/contrib/Descriptions/lme4.html.

Berent, Iris (2008). Are phonological representations of printed and spoken language isomorphic? Evidence from the restrictions on unattested onsets. *Journal of Experimental Psychology: Human Perception & Performance, 34*, 1288-1304.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno, & Paul Smolensky (2008). Language universals in human brains. *Proceedings of the National Academy of Science*, 105(14), 5321-5325.

Berent, Iris, Donca Steriade, Tracy Lennertz, & Vered Vaknin (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104(3), 591-630.

Blevins, Juliette (2004). *Evolutionary Phonology: The emergence of sound patterns.* Cambridge: Cambridge University Press.

Chomsky, Noam & Morris Halle (1968). *The Sound Pattern of English*. Harper and Row, New York.

Clements, George N. (1988). The sonority cycle and syllable organization. In Dresher et al. (eds.). *Phonologica 1988*. Cambridge: Cambridge U. Press.
Coady, Jeffrey A. & Julia L. Evans (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language Communication Disorders*, 43(1), 1-40.

Coetzee, Andries W. (2008). Grammaticality and Ungrammaticality in Phonology. *Language*, 84(2), 218-257.

Coleman, John S. & Pierrehumbert, Janet B. (1997). Stochastic phonological grammars and acceptability. *Computational Phonology*, 3, 49-56.

Coppieters, Rene (1987). Competence differences between native and near-native speakers. *Language*, 63(3), 544-573.

Daland, Robert & Pierrehumbert, Janet B. (2011). Learning diphone-based segmentation. *Cognitive Science*, 35(1), 119-155.

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

Davidson, Lisa (2007). The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology*, 24(2), 261-286.

Davidson, Lisa (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* 38:2, 272-288.

Della Pietra, Stephen, Vincent J. Della Pietra, & John D. Lafferty (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 380-393.

Duanmu, San (2000). *The Phonology of Standard Chinese*. Oxford University Press.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier & Jacques Mehler (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance 25*(6), 1568-1578.

Edwards, Jan, Mary E. Beckman, & Benjamin Munson (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47, 421-436.

Frisch, Stefan A., Michael B, Broe, & Janet B. Pierrehumbert (1997). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22, 179-228.

Frisch, Stefan, and Zawaydeh, Bushra (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77, 91-106.

Gale, William A., and Geoffrey Sampson (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics 2*(3), 217-237.

Goldrick, Matthew. (in press). Utilizing psychological realism to advance phonological theory. In J. Goldsmith, J. Riggle, & A. Yu (Eds.) *Handbook of phonological theory* (2nd edition). Blackwell.

Goldsmith, John & Jason Riggle (to appear). Information Theoretic Approaches to Phonological Structure: The Case of Vowel Harmony. *Natural Language and Linguistic Theory*.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120.

Hay, J., Pierrehumbert, J., & Beckman, M. (2003). Speech perception, well-formedness,and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.) *Phonetic Interpretation: Papers in Laboratory Phonology VI* (pp. 58-74). Cambridge University Press.

Hayes, Bruce (in press). Interpreting sonority-projection experiments: the role of phonotactic modeling. To appear in *Proceedings of 17th International Congress of Phonetic Sciences, Hong Kong*.

Hayes, Bruce & Donca Steriade (2004). Introduction: The phonetic bases of phonological markedness. In B. Hayes, R. Kirchner, & D. Steriade (eds.) *Phonetically-based phonology*. Cambridge: Cambridge University Press.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39, 379-440.

Hockett, Charles F. (1947). Peiping phonology. *Journal of the American Oriental Society 67*, 253-267. Reprinted in Martin Joos (ed.), *Readings in Linguistics I: The Development of Descriptive Linguistics in America 1925-56 (4th Ed.)*, 217-228. University of Chicago Press: Chicago.

Hooper, J. B. (1976). *An introduction to natural generative phonology*. New York: Academic Press.

Jespersen, Otto (1904). *Lehrbuch der Phonetik*. Leipzig and Berlin.

Jurafsky, Daniel & James H. Martin (2009). *Speech Processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd edition).* New Jersey: Prentice Hall.

Joseph, John E. (1995). Natural grammar, arbitrary lexicon: An enduring parallel in the history of linguistic thought. *Language & Communication* 15(3), 213-225.

Kager, Rene & Joe Pater (under revision). Phonotactics as phonology: Knowledge of a complex restriction in Dutch.

Kawahara, Shigeto (2008). Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics*, 17, 317-330.

Kawahara, Shigeto (ms). Modes of phonological judgement. ROA

Koo, Hahn & Jennifer Cole (2006). On learnability and naturalness as constraints on phonological grammar. In Antonis Botinis (ed.) *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*, 174-177. Athens: University of Athens.

Lee, Yongsung (1994). Onset Analysis of Korean On-Glides. In Young-Key Kim-Renaud (ed.) *Theoretical Issues in Korean Linguistics* (pp. 133-156). CSLI Publications: Stanford.

Legendre, Géraldine, Yoshiro Miyata, & Paul Smolensky (1990). Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: an application. *COGSCI 1990*, 884–891.

Manning, Chris & Hinrich Schütze (1999). *Foundations of statistical natural language processing*. MIT Press: Cambridge, MA.

McFadden, Daniel. 1974. Conditional logit analysis of qualitative choice behaviour. In Zarembka, Paul (ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

Parker, Stephen G. (2002). Quantifying the sonority hierarchy. University of Massachusetts, Amherst Doctoral Dissertation, Paper AAI3056268. http://scholarworks.umass.edu/dissertations/AAI3056268

Pater, Joe (2009). Weighted constraints in generative linguistics. *Cognitive Science*, 33, 999-1035.

Pierrehumbert, Janet B. (2006) The Statistical Basis of an Unnatural Alternation, in L. Goldstein, D.H. Whalen, and C. Best (eds), *Laboratory Phonology VIII, Varieties of Phonological Competence*. Mouton de Gruyter, Berlin, 81-107.

Prince, Alan and Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar.* Technical Report 2, Rutgers University Center for Cognitive Science.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker (2010). Harmonic Grammar with linear programming: From linear systems to linguistic typology. *Phonology* 27, 77-117.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*. Vienna: Austria.

Ren, Jie, Liqun Gao, & James L. Morgan (2010). Mandarin speakers' knowledge of the sonority sequencing principle. *Presented at the 20th Colloquium on Generative Grammar at the Universitat Pompeu Fabra, Barcelona, March 18-20.*

Selkirk, Elizabeth O. (1982). The syllable. In Harry van der Hulst and Norval Smith (eds.) *The Structure of Phonological Representations, Part II*. Dordrecht: Foris. Pp. 337-383.

Selkirk, Elizabeth (1984). On the major class features and syllable theory. In M. Aronoff & R. T.Oehrle (Eds.) *Language sound structure: Studies in phonology presented to Morris Halleby his teacher and students*, (pp. 107-136). Cambridge, MA, London: The MIT Press.

Sievers, Eduard (1881). Grundzuge der Phonetik. Breitkopf und Hartel, Leipzig.

Smolensky, Paul, and Géraldine Legendre (2006). *The harmonic mind: from neural computation to Optimality-theoretic grammar.* Cambridge: MIT Press.

Steriade, Donca (1982). Greek Prosodies and the Nature of Syllabification. PhD dissertation, MIT, Cambridge, Massachusetts.

Vitevitch, M.S. & Paul A. Luce (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, 36, 481-487.

Wilson, Colin and Lisa Davidson (in press). Bayesian analysis of non-native cluster production. In Proceedings of NELS 40, MIT, Cambridge, MA.

Wright, Richard (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (eds.) *Phonetically-based phonology*. Cambridge: Cambridge University Press.