Gradient phonological acceptability as a grammatical effect

Adam Albright
MIT

Draft: January 2007 (Comments welcome)

Abstract

Phonological judgments are often gradient: *blick* > ?*bwick* > **bnick*. The theoretical interpretation of gradient acceptability remains controversial, however, with some authors maintaining that it is a performance/task effect based on similarity to the lexicon (*neighborhood effects*), and others attributing it to a probabilistic grammar regulating possible sequences (*phonotactics*). In a study that directly compared the predictions of similarity-based and sequential models against experimental ratings of non-words, Bailey and Hahn (2001) argued that both types of knowledge are needed, though the relative contribution of sequential models was quite small. In this paper, additional phonotactic models are considered, including the widely used positional phonotactic probability model of Vitevitch and Luce (2004), and a model based on phonological features and natural classes. The performance of all models is tested against Bailey and Hahn's data and against data from Albright and Hayes (2003). The results show that probabilistic phonotactic models do not play a minor role; in fact, they may actually account for the bulk of gradient phonological acceptability judgments.

1 Introduction

When native speakers are asked to judge made-up (nonce) words, their intuitions are rarely all-or-nothing. In the usual case, novel items fall along a gradient cline of acceptability.[1] Intermediate levels of acceptability are illustrated for a selection of novel pseudo-English words in (1). These range from words like [stɪn], which speakers tend to deem quite plausible as a possible word of English, to [ʃœb], which is typically judged to be implausible or outlandish.

[1] I use the term gradient *acceptability* rather than gradient *wordlikeness* or *phonotactics*, since those terms presuppose at least implicitly a particular underlying mechanism.

(1) "How good would . . . be as a word of English?"

Best	*stin* [stɪn] , *mip* [mɪp]
	blick [blɪk], *skell* [skɛl]
Intermediate	*blafe* [bleɪf], *smy* [smaɪ]
	bwip [bwɪp], *smum* [smʌm]
	dlap [dlæp], *mrock* [mrak]
Worst	*bzarshk* [bzarʃk], *shöb* [ʃœb]

Numerous studies have demonstrated a relation between gradient acceptability and the statistics of the lexicon, using data from different domains and different experimental methodologies. The relation between behavior on nonce words and the number of similar existing words has been explored extensively in the area of morphological productivity (Bybee 1985; Anshen and Aronoff 1988; Bybee 1995; Plag 1999; Bauer 2001; Albright 2002a; Albright and Hayes 2003; and many others), and in the area of morphophonological alternations (Berko 1958; Bybee and Pardo 1981; Eddington 1996; Zuraw 2000; Bybee 2001; Pierrehumbert 2002; Ernestus and Baayen 2003). Of more immediate relevance to the current study, gradient phonological acceptability has also been shown to relate to lexical statistics (Greenberg and Jenkins 1964; Scholes 1966; Ohala and Ohala 1986; Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000; Bailey and Hahn 2001; Hay, Pierrehumbert, and Beckman 2004). The strategy for demonstrating a lexicostatistic effect is to quantify differences between novel items, typically in terms of neighborhood density (Greenberg and Jenkins 1964; Coltheart, Davelaar, Jonasson, and Besner 1977; Luce 1986; Newman, Sawusch, and Luce 1997; Luce and Pisoni 1998) or biphone probabilities (Vitevitch, Luce, Charles-Luce, and Kemmerer 1997; Vitevitch and Luce 1998), and show that there is a positive correlation between speakers' behavior and the number of similar existing words.

Such work has made it abundantly clear that speakers' behavior is guided in some fashion by knowledge of lexical statistics. The exact mechanism by which this happens is considerably more controversial, however. As Bailey and Hahn (2001) point out, lexical statistics such as neighborhood density and biphone probability are often highly correlated with one another, so special care is required to tease apart any independent predictions that they might have (e.g., Vitevitch, Luce, Pisoni, and Auer 1999). More typically, gradient effects are simply assumed to arise either from neighborhood similarity to the lexicon or to a stochastic grammar, according to the authors' theoretical predisposition. According to one prevalent point of view, gradient phonological effects do not have anything to do with grammar, but rather arise as part of the task of processing and judging novel sequences. This point of view is held by those who deny grammar altogether and attribute all aspects of linguistic behavior to properties of language use (Bybee 2001). It is also the view held by linguists who view grammar as categorical, and fundamentally unable to derive gradient effects (for discussion, see Schütze 2005). For proponents of this view, grammar defines what is a *possible* word, whereas acceptability tasks require speakers to judge how *probable* the novel item would be as a word of the language. This calculation may invoke a variety of different non-grammatical calculations, such as assessing the number of similar words that are activated during the attempt to carry out lexical access, or performing some more elaborate form of analogical reasoning. For scholars of both stripes, gradient acceptability is seen as a performance effect, accessing types of

knowledge that go beyond what is encoded in the grammar (which, if it exists, is categorical and non-probabilistic).

These views can be contrasted with one in which grammar itself is seen as probabilistic and gradient. Under such a view, grammaticality is not merely a matter of set membership (grammatical vs. ungrammatical), but reflects a probability distribution over sequences of segments. Under this view grammatical knowledge of probabilities is informed by (or even completely determined by) statistics of the lexicon, but the grammar represents a learned abstraction that is independent from the lexicon itself. A stochastic grammar can assess the probability of strings regardless of their existence as actual words, and gradient acceptability judgments directly mirror gradient grammaticality intuitions (Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000; Albright and Hayes 2003; Hammond 2004; Hay and Baayen 2005; Hayes and Wilson pear).Of course, it is also possible that gradient acceptability judgments rely on both a stochastic grammar and also lexical comparison; (Bailey and Hahn 2001) explicitly claim that both kinds of knowledge are needed.

How do we distinguish among these possibilities? Simply demonstrating that there are gradient effects clearly cannot settle the issue, since the disagreement is not about the existence of such effects, but rather, about their proper interpretation. Unfortunately, in most cases, lexical vs. grammatical accounts of gradient experimental results are not tested in parallel and weighed against one another—and indeed, doing so can be difficult, given their high degree of intercorrelation. Bailey and Hahn (2001) present side-by-side comparisons of both lexical and sequential models on the same batch of gradient acceptability judgments, finding independent effects of each, but with a rather greater contribution of lexical neighborhood effects. This result is ambiguous from the point of view of motivating a probabilistic phonological grammar. On the one hand, the authors did find a significant contribution of gradient knowledge of possible sound sequences, contrary to their initial hypothesis. On the other hand, the contribution was quite small, adding only single-digit gains (6%—9%) to their overall model. The goal of the current study is to examine in greater detail the evidence for probabilistic phonotactic knowledge, both to determine whether it is truly necessary in explaining gradient phonological acceptability, and also to situate the finding that it is relatively unimportant compared to knowledge of lexical neighborhoods. The form of the study is directly inspired by Bailey and Hahn's comparison, and is in fact partly a direct replication of their Experiment 2. In section 2, I will consider several computationally implemented models of gradient acceptability, including similarity-based neighborhood models (capturing the role of on-line lexical access) and sequence-based phonotactic models (reflecting the role of probabilistic grammars). The performance of these models is tested in section 3 against two data sets: one from Bailey and Hahn's own study, and one collected by Albright and Hayes (2003). The two data sets show strikingly different results: whereas item-by-item differences in the Bailey and Hahn data are largely ambiguous (though slightly better explained as neighborhood effects), the differences in the Albright and Hayes data diagnose a primarily phonotactic effect. In section 4, I consider possible reasons for this discrepancy, concluding that currently available data sets most likely underestimate the importance of phonotactic knowledge in assessing gradient phonological well-formedness. It is argued that the role of probabilistic phonological knowledge cannot be dismissed, and that gradient acceptability judgments are at their core grammaticality judgments.

2 Lexical vs. combinatorial models of gradience

Bailey and Hahn (2001) distinguish between two fundamentally different types of models of how speakers judge the probability of novel words. The first is an *analogical model*, which makes use of solely lexical knowledge. Such a model is based on the undeniable premise that that speakers have knowledge of existing words, and that hearing a novel word activates a set of real words in the process of attempting to recognize the word. Intuitively, the greater the number of words that are activated, and the greater their similarity to the novel word, the more lexical support the word receives, and the higher its score. By contrast, a *combinatorial model* makes use of knowledge of combinatorial phonotactic possibilities of different sounds in the language. This is based on the premise that speakers are able to decompose words into consituent parts (e.g., segments), and assess the likelihood of combining those parts in that order. In this type of model, novel words that consist of probable combinations of sounds receive correspondingly higher scores than those consisting of improbable/"illegal" sequences.[2] It is not at all implausible to think that speakers have both lexical and phonotactic knowledge, as is commonly assumed both by generative phonologists (Chomsky and Halle 1968) and by psycholinguists (e.g., Vitevitch and Luce 2005). It is nonetheless useful to investigate the predictions of extreme "endpoint models" that make use of lexical or phonotactic knowledge alone, since such models make maximally distinct predictions about the factors that should influence gradient acceptability. In particular, if gradient acceptability depends on consulting the lexicon, then we should effects of factors that are known to play a role in lexical access, such as lexical (token) frequency, neighborhood density, and so on. If, on the other hand, gradient acceptability is a result of a probabilistic grammar, we should expect to see a role for factors that are known to play a role in grammatical descriptions, such as natural classes, phonological markedness, the type frequency of a generalization, the number of exceptions it has, and so on. The strategy of this study, therefore, is to contrast models that use either lexical or sequential knowledge to predict the acceptability of novel sequences, and test to what extent their ability to use different sources of information helps their performance in modeling experimentally obtained ratings.

2.1 Lexical models of gradient acceptability

The most widely used technique for estimating the degree of lexical support for a novel item is to calculate its NEIGHBORHOOD DENSITY, defined as the number of words that differ from the target word by *n* changes (substitutions, additions, or deletions) (Luce 1986). For example, the nonce word *droff* [draf] has neighbors such as *trough*, *prof*, *drop*, *dross*, and *doss* (the exact set depends, naturally, on the dialect of English). Neighborhood density calculated in this way has been shown to predict a wide variety of effects, including lexical decision times (Luce and Pisoni 1998), phoneme identification (Newman, Sawusch, and Luce 1997), mishearings (Vitevitch 2002), and acceptability of novel words (Greenberg and Jenkins 1964; Ohala and Ohala 1986). The

[2] A categorical model of grammar is a special case of combinatorial grammar, in which sequences are assigned probabilities of 1 (grammatical) or 0 (ungrammatical).

classic definition of neighbors is widely acknowledged to be only a crude estimate of similarity, since it does not take into account segmental similarity (Frisch 1996; Hahn and Bailey 2005), the perceptual difference between changing vs. removing a segment, or the relative perceptual salience of segments in different positions within the word (Vitz and Winkler 1973; Cutler and Norris 1988). As Bailey and Hahn (2001) point out, this is an issue especially for modeling behavior on non-words, since even relatively ordinary non-words often have no single-modification neighbors—e.g., *drusp* [drʌsp], *stolf* [stalf], and *zinth* [zɪnθ] have a neighborhood density of zero under the traditional definition, putting them on a par with words like *shöb* [ʃœb] or *bzarshk* [bzarʃk]. An adequate model of lexical support must be able to take into account the fact that although words like *drusp* have no immediate neighbors, they are nonetheless quite similar to a number of words (*trust*, *dusk*, *rusk*, *truss*, *dust*, etc.). This requires a model that can count words more than one change away, while at the same time paying attention to the severity of different substitutions.

In order to achieve this, Bailey and Hahn propose the GENERALIZED NEIGHBORHOOD MODEL (GNM), an adaptation of the GENERALIZED CONTEXT MODEL (GCM; Nosofsky 1986, 1990). The GCM is a similarity-based classification model that categorizes novel exemplars based on their aggregate similarity to sets of existing exemplars. Bailey and Hahn's GNM is adaptation specifically designed to quantify lexical neighborhood effects. In the GNM, the set of existing exemplars is defined as the lexicon of known words, and the degree of lexical support for a novel word is proportional to a weighted sum of the perceptual similarities of the novel word *i* to each existing word *w*. The formal definition, given in (2), uses an exponential function to convert the transformational distance between *i* and *w* ($d_{i,w}$) to an estimate of perceived psychological similarity.

(2) Definition of the GNM:

$$\text{Support for item } i = \sum \text{weight}_w \times e^{(-d_{i,w}/s)}, \text{ where}$$

- weight_w = a function of the token frequency of word *w*
- $d_{i,w}$ = psychological distance between nonce item *i* and existing word *w*
- s = sensitivity, a parameter that controls the magnitude of the advantage that very similar neighbors have in determining the outcome
- $e \approx 2.71828$

The distance $d_{i,w}$ between novel and existing words is calculated by finding their minimum string edit distance (Kruskal 1983/1999, chap. 1; Jurafsky and Martin 2000, §5.6). This involves finding the optimal alignment between the segments of *i* and those of *w*, combining the best pairing of phonetically similar segments and the fewest possible insertions and deletions. Phonetic similarity between potentially corresponding segments is assessed with a metric based on shared vs. unshared natural classes, following Frisch, Pierrehumbert, and Broe (2004). The cost of insertions and deletions is a free parameter of the model, whose optimal value must be found by fitting.[3] For

[3]Ultimately, it would be desirable to weight similarity according to location of mismatches (syllable position, word-initial vs. medial, etc.), and prosodic factors like stress. However, since all of the words modeled in §3 are monosyllabic, reasonable results can be obtained even without a prosodically sensitive model.

further details of the GNM, see Bailey and Hahn (2001); for other linguistic applications of the GCM, see Johnson (1997), Nakisa, Plunkett, and Hahn (1997), Albright and Hayes (2003), and Wilson (2006).

The concrete result of this model is that nonce words receive greater lexical support from existing words with lesser phonetic dissimilarity. This is illustrated for *zin* [zɪn] vs. *snulp* [snʌlp] in (3)–(4); in each case, the five most similar analogs are shown, along with their predicted perceptual dissimilarity (in arbitrary units).

(3) Lexical support for *zin* [zɪn]: many similar words

Existing word		Dissimilarity
zen	[zɛn]	0.609
sin	[sɪn]	0.613
din	[dɪn]	0.649
in	[ɪn]	0.700
zing	[zɪŋ]	0.720

(4) Lexical support for *snulp* [snʌlp]: more distant, and drops off quickly

Existing word		Dissimilarity
snub	[snʌb]	1.260
sulk	[sʌlk]	1.271
slump	[slʌmp]	1.283
snuff	[snʌf]	1.367
null	[nʌl]	1.400

Finally, it is intuitively possible that in addition to phonetic similarity, lexical frequency may also play a role in determining the relative contribution of existing words. In a canonical exemplar model, every token of an existing word would add its own support, giving greater weight to high frequency words. However, it has been hypothesized that speakers may treat very high frequency words as idiosyncratic or autonomous entities, with the result that mid-frequency words may actually have the greatest say in determining the outcome for novel items (Bybee 1995). For this reason, Bailey and Hahn implement the effect of frequency (*weight*$_w$ in (2)) as a quadratic function of the log token frequency of word *w*, which is able to capture both monotonic and non-monotonic frequency effects—and indeed, they find that the optimal function is one which gives medium frequency words the greatest say. As with the other parameters of the model, the exact shape of the frequency effect must be determined post hoc by fitting.

The GNM is by no means the only model that attempts to assess the degree of overlap that an incoming word has to the set of existing words—other popular models include Cohort (Marslen-Wilson and Welsh 1978), TRACE (McLelland and Elman 1986) and Shortlist (Norris 1994). Among these, the GNM is one of the most flexible in its ability to accommodate different theories of phonetic similarity and lexical frequency. The large number of free parameters can be a liability, in that a good deal of post hoc fitting is required; however, for present purposes it is also

an asset, since it allows us to explore independently the contribution of various factors such as token frequency, phonetic similarity, and so on.

2.2 Phonotactic models of gradience

In contrast to lexical/analogical models, which assess whole-word similarity to existing items, phonotactic models evaluate local substrings of words. Every theory of phonological grammar provides a mechanism for distinguishing grammatical from ungrammatical combinations of sounds or classes of sounds, usually described in terms of phonological features. For example, both the rule-based approach of the *Sound Pattern of English* (Chomsky and Halle 1968) and the constraint-based approach of Optimality Theory (Prince and Smolensky 1993/2004) contain mechanisms for targeting structural descriptions, stated in terms of matrices of phonological features. Grammars may restrict co-occurrence of feature values within a single segment, as in the ban on simultaneous frontness and roundness in English vowels ((5a)), or they may restrict combinations in nearby segments, as in the ban on the segment *n* followed by tautosyllabic obstruents with a different place of articulation ((5b)).[4]

(5) a. No front rounded vowels: $*\begin{bmatrix} -\text{back} \\ +\text{round} \end{bmatrix}$

b. Nasal place assimilation: $*\begin{bmatrix} +\text{nas} \\ +\text{coronal} \end{bmatrix}\begin{bmatrix} -\text{son} \\ -\text{coronal} \end{bmatrix}]_{\sigma}$

A simple and widely used class of computational models for calculating the probability of strings of adjacent elements are N-GRAM MODELS (Jurafsky and Martin 2000, chap. 6), which keep track of the probabilities of substrings *n* elements long. For example, a bigram model of phonology would estimate the probability of a string of phones *abcd* by considering the probability of each two phone substring (*#a*, *ab*, *bc*, *cd*, *d#*). Typically, the probability of a substring *ab* is defined as the conditional probability of *b* given *a* (i.e., the TRANSITIONAL PROBABILITY $P(b|a)$).

(6) Bigram transitional probability:

$$\text{Probability of } b \text{ given } a = \frac{\text{Number of times } ab \text{ occurs in the corpus}}{\text{Total number of times } a \text{ appears before anything}}$$

Various strategies have been employed in the literature to combine local transitional probabilities into an overall score for an entire word *abcd*. The approach most in keeping with the use of *n*-gram models in syntax is to calculate the joint probability of all local *n*-segment sequences

[4]The proper characterization of locality has been a major focus of theoretical phonology. In this paper, I will consider restrictions only among strictly adjacent elements; for an approach to discovering constraints on non-local (long-distance) elements, see Albright and Hayes (2006).

co-occurring in the same word ((7a)). In this model, a single unattested sequence within the word has the consequence of driving the entire score to zero, not unlike the idea in Optimality Theory that a single violation can prove fatal to a candidate. Coleman and Pierrehumbert (1997) note that this prediction does not always seem to be true, and propose instead a metric that allows scores from different parts of the word to trade off against one another. The simplest modification to an *n*-gram model that has this property is to consider the *average transitional probability*, as in (7b). In practice, most studies in the literature that have attempted to examine or control for bigram probability have used average probabilities (Vitevitch, Luce, Charles-Luce, and Kemmerer 1997; Vitevitch and Luce 1998; Bailey and Hahn 2001).

(7) Probability of a sequence *abcd*

a. Joint transitional probability

$$\text{P}(abcd) = \text{P}(\text{initial } a) \times \text{P}(b|a) \times \text{P}(c|b) \times \text{P}(d|c) \times \text{P}(\text{ending after } d)$$

b. Average transitional probability

$$\text{P}(abcd) = \text{Average}(\text{P}(\text{initial } a), \text{P}(b|a), \text{P}(c|b), \text{P}(d|c), \text{P}(\text{ending after } d))$$

An additional modification, proposed by Vitevitch and Luce (1998, 2004) is to calculate biphone probabilities separately for each position in the word (initial, second position, third position, etc.). This is related to, but implementationally quite distinct from, the idea that the probability of sequences may be sensitive to prosodic position. More refined metrics taking into account stress and syllable position have also provided significant improvement over simple frequency scores (Coleman and Pierrehumbert 1997; Treiman, Kessler, Knewasser, Tincoff, and Bowman 2000; Frisch, Large, and Pisoni 2000; Hay, Pierrehumbert, and Beckman 2004). What all of these metrics have in common is that they impose structure on novel words, parsing the string into subconstituents and evaluating their likelihood of co-occurrence.

Biphone probabilities in one form or another have been shown to correlate with a wide variety of phenomena, including phoneme identification (Pitt and McQueen 1998), repetition speed in shadowing tasks (Vitevitch, Luce, Charles-Luce, and Kemmerer 1997; Vitevitch and Luce 1998; Vitevitch and Luce 2005), response time in same/different tasks (Vitevitch and Luce 1999), looking times in 9-month olds (Jusczyk, Luce, and Charles-Luce 1994), and most relevant for present purposes, wordlikeness judgments (Vitevitch, Luce, Charles-Luce, and Kemmerer 1997; Bailey and Hahn 2001).In theory, one might consider higher order models that take into account the preceding two segments (trigram model) or more, but in practice, there is a trade-off between adopting a finer-grained model and having sufficient data available about each *n*-segment sequence. For example, the novel word *dresp* [drɛsp] contains the sequence [ɛsp], which is unattested as a syllable rhyme in English.[5] More generally, for larger sequence lengths, the number of possible *n*-grams grows exponentially, meaning a much larger corpus is needed to estimate their frequencies. In the present case the corpus is the lexicon of attested words, which is of a limited size. Hence,

[5]The sequence [ɛsp] is attested across syllable boundaries, in words like *desperate*, *trespass*, *espionage*, *cesspool*, *despot*, and *desperado*.

it is not possible to get more data about the goodness of [ɛsp] by simply collecting more words. Numerous techniques have been proposed to avoid underestimating the goodness of accidentally unattested sequences, such as reserving some probability mass for unseen sequences (smoothing) and combining information from shorter and longer values of *n* (backoff); see Jurafsky and Martin (2000, chap. 6) for an overview. A strategy suggested by decades of work in phonological theory is to reason about the well-formedness of underattested sequences based on natural classes—that is, based on the occurrence of other, featurally similar sequences.

The insight of generalization based on natural classes is the following: although words ending in [ɛsp] are unattested in English, several minimally different sequences are in fact well attested, such as [ɪsp] (*crisp*, *wisp*, *lisp*), [æsp] (*clasp*, *rasp*, *asp*), [ɛst] (*best*, *west*, *rest*), [ɛsk] (*desk*) and so on. Taken together, these words strongly suggest that English allows sequences of lax vowels followed by *s* and a voiceless stop (*p*, *t*, *k*)—a conclusion that has been made explicitly in phonological analyses of English (e.g., Hammond 1999, p. 115). In order to discover this, the learner must be able to consider not just co-occurrences of segments as atomic units, but also of feature combinations that define classes of segments.

One proposal for how speakers compare sequences to extract more general patterns of natural classes is the MINIMAL GENERALIZATION approach (Albright and Hayes 2002, 2003). The idea of this approach is that learners discover grammatical patterns by abstracting over pairs of strings, extracting the features that they have in common. The procedure is minimal in the sense that all shared feature values are retained; the algorithm thus conservatively avoids generalization to unseen feature values. For example, comparison of [ɪsp] and [æsp] yields a generalization covering all front lax vowels ((8a)), while further comparison with [ɛsk] extends the generalization to all front lax vowel + *s* + voiceless stop sequences ((8b)).[6]

(8) Minimal Generalization

a.		ɪ	s	p
	+	æ	s	p
	→	[−back, −round, −tense]	s	p
b.	+	ɛ	s	k
	→	[−back, −round, −tense]	s	[−sonorant, −contin., −voice]

[6]The precise inferences that are available depend intimately on the set of features that is assumed. For example, if [coronal] is treated as a binary feature with both [+coronal] and [−coronal] values, then *p* and *k* share a [−coronal] specification and comparison of [æsp] and [ɛsk] would not extend to *Vst* sequences. In general, bivalent (equipollent) feature definitions limit abstraction, by creating more "negatively specified" natural classes such as [−coronal] and preventing generalization to unseen positive values. For present purposes I will assume privative place features (i.e., '+' values with no corresponding '−' value).

In the example in (8), the sequences under comparison differ rather minimally in their feature values, and the resulting generalizations seem intuitively rather well-supported. Not all comparisons yield equally illuminating abstractions, however. Consider, for example, the comparison of *asp* [æsp] and *boa* [boʊə] in (9). Here, the initial segments of both strings are voiced, but the remaining segments have little or nothing in common featurally. The resulting abstraction, therefore, is simply one in which voiced segments can be followed by two additional segments.

(9) A minimal but sweeping generalization

	æ	s	p
+	b	oʊ	ə
→	[+voi]	*seg*	*seg*

The pattern in (9) is well attested in English: voiced segments are very often followed by at least two additional segments, and segments are frequently preceded by voiced segments two to their left. Taken seriously, however, this pattern makes undesired predictions: it leads us to expect that *any* combination of three segments in which the first is voiced should be possible. This leads to the potentially fatal prediction that the nonce word *bzarshk* [bzarʃk] should be very acceptable, since the sequences [bza], [arʃ], and [rʃk] get a good deal of support from attested '[+voi] *seg seg*' sequences.

The challenge, then, is to find a way to count over natural classes so that [æsp] and [ɪsp] provide very strong support for [ɛsp], while [æsp] and [boə] provide little or no support for [bza] or [rʃk]. Intuitively, the inductive leap from [ɪsp] and [æsp] to [ɛsp] is quite small; in fact, the comparison makes [ɛsp] seem practically inevitable. This is no doubt do at least in part to the fact that the resulting abstraction, '[−back,−round,−tense] *sp*' is extremely specific—all we need to specify in order to get [ɛsp] is the vowel height. In order to get [ɛsp] from '[+voice] *seg seg*', on the other hand, we must fill in very many feature values. The former abstraction specifically "speaks to" [ɛsp] in a way that the latter does not.

To see the same point from a different perspective, consider the situation of a learner who has heard only the two data points [ɪsp] and [æsp]. The '[−back,−round,−tense] *sp*' pattern describes a very small set of possible sequences (namely, [ɪsp], [ɛsp] and [æsp]). If we were to adopt the hypothesis that such sequences are legal without making any further assumptions, we would expect the probability of encountering any one of them at random to be 1/3. Thus, we expect that there is a very high chance that we might encounter [ɛsp] as well. By contrast, if we adopt the hypothesis that English allows any '[+voi] *seg seg*' sequence, then the chance of getting attested [ɪsp] or [æsp] at random from among all logically possible [+voi] *seg seg* combinations is tiny. Although both patterns can describe the existing data, the more specific characterization makes it much less of a coincidence that [ɪsp] and [æsp] were the two words that we happened to actually receive in the training data, and for the same reason, makes it much more likely that [ɛsp] would also be a word of the language. The goal of the learner can be defined as finding the set of statements that characterize the data as tightly as possible, under the assumption that words conform to certain shapes because the grammar forces them to, and not out of sheer coincidence. This is related to the

principle of MAXIMUM LIKELIHOOD ESTIMATION (MLE), which seeks to find the description that makes the training data as likely as possible.[7] It is also related to proposals within Optimality Theory for seeking the most restrictive possible grammar (Prince and Tesar 2004; Hayes 2004; Tessier 2006), as a way of obeying the subset principle (Berwick 1986).

In order to test the usefulness of this principle, a model was constructed that evaluates combinations of natural classes based not only on their rate of attestation, but also on their specificity. If the *n*-gram probability of a sequence of two literal segments *ab* is defined as in (10a), then the probability of the segments *ab* as members of natural classes *A*, *B* respectively can be defined as in (10b).

(10) a. Probability of the sequence $ab = \frac{\text{Number of times } ab \text{ occurs in corpus}}{\text{Total number of two-item sequences}}$

b. Probability[8] of the sequence *ab*, where $a \in$ class *A*, $b \in$ class *B*

$\propto \frac{\text{Number of times } AB \text{ occurs in corpus}}{\text{Total number of two-item sequences}} \times \text{P(choosing } a \text{ from } A) \times \text{P(choosing } b \text{ from } B)$

As mentioned above, one intuitive way to define the probability of selecting a particular member of a natural class is to assume that each member of the class is equally probable, so that the probability of choosing any member at random is inversely proportional to the size of the class ((11a)). A slightly more sophisticated model would take into account the fact that different segments independently have different frequencies, and to weight the probabilities of choosing members of a natural class accordingly ((11b)). For the simulations reported below I assume the simpler definition, which is not sensitive to token frequency.

(11) Two definitions of instantiation costs

a. Simple: $\text{P(member } a|\text{class } A) = \frac{1}{\text{Size of } A \text{ (i.e., number of members)}}$

b. Weighted: $\text{P(member } a|\text{class } A) = \frac{\text{Token frequency of } a}{\sum \text{Token frequency of all members of } A}$

[7] For an MLE-based approach to *n*-gram models that refer to classes of items, see Saul and Pereira (1997). Unfortunately, *n*-gram models based on classes of elements in syntax tend to assume that each item belongs to ideally one, or exceptionally a few classes (‘*tree*’ is a verb, ‘*of*’ is a preposition, etc.). For phonological applications, we need to consider each segment as a member of many classes simultaneously, and even a single instance of a segment may be best characterized in different terms with respect to what occurs before it vs. after it. Ultimately, we seek a learning algorithm that incorporates the principle of MLE, without the assumptions about class structure that existing class-based *n*-gram models typically make.

[8] Since natural classes in phonology are characteristically overlapping and each segment belongs to many natural classes, the equation in (10a) will not yield values that sum to 1 over all segments and natural classes; a normalization term would be needed to convert these scores to true probabilities.

Since each segment belongs to multiple natural classes ([ɛ] is a vowel, a front vowel, a front lax vowel, a sonorant, a voiced segment, etc.), there are many ways to parse a string of segments into sequences of natural classes. For example, as noted above, [ɛsp] could be considered a member of the trigram '[−back,−round,−tense] *sp*', or of the trigram '[+voi] *seg seg*'. The probability of the sequence [ɛsp] as a member of the trigram '[−back,−round,−tense] *sp*' depends not only on the probability of '[−back,−round,−tense] *sp*' combinations, but also on the probability of instantiating the [−back, −round, −tense] class as an [ɛ]. Although the frequency of '[−back,−round,−tense] *sp*' sequences is not especially high in English, the probability of instantiating [−back,−round,−tense] as [ɛ] is quite high (= 1/3, given the simple definition in (11a)), yielding a relatively high overall predicted probability. The probability of [ɛsp] as an instantiation of the trigram '[+voice] *seg seg*', on the other hand, relies not only on the frequency of '[+voice] *seg seg*' combinations (which is relatively high) but also on the probability of [ɛ] as a voiced segment (≈ 1/35), as well as *s* and *p* among the entire set of possible segments (≈ 1/44 each). This yields an instantiation cost of $\approx 1.48 \times 10^{-5}$, and a correspondingly low predicted probability. It seems reasonable to assume that among the many possible ways of parsing a given sequence into natural classes, the one with the highest probability is selected (for a similar assumption, see Coleman and Pierrehumbert 1997 and Albright and Hayes 2002).

To summarize, I have sketched here a method of evaluating sequences of natural classes in order to determine which combinations of feature values are supported by the training data. The model takes into account not only the frequency of the combinations, but also their specificity, combined according to the definition in (10b). Taken together, the need to maximize high *n*-gram probabilities and minimize instantiation costs should drive the model to seek characterizations that involve frequent combinations of features while remaining as specific as possible. The result is a stochastic grammar that permits novel sequences to be assigned likelihood scores, by attempting to parse them into well-supported combinations of natural classes.

3 Testing the models

In order to evaluate the usefulness different types of models, it is useful to be able to compare their performance side-by-side on the same set of data. In this section, I present the results of simulations using both lexical and phonotactic models to predict native speaker ratings of novel English words, from two previous nonce word studies: Bailey and Hahn (2001) and Albright and Hayes (2003). This comparison is in part a replication of Bailey and Hahn's study, since it includes some of the same models tested on the same data set. The purpose of this replication is to allow a direct comparison with data from other studies, and to compare models that Bailey and Hahn did not consider (in particular, the model based on natural classes, presented in the preceding section).

In all, six models of gradient well-formedness were compared. The first two were lexical models: one that assigned scores based on the traditional definition of neighborhood density ("differ by one change"), and the refined model proposed by Bailey and Hahn, the Generalized Neighborhood

Model (GNM). In addition, four phonotactic models were considered: joint bigram and trigram transitional probability ((7) above),[9] position-dependent bigram probability (Vitevitch and Luce 2004)[10], and the model that generalizes based on natural classes, sketched above.

Except for the Vitevitch and Luce model, which uses its own smaller dictionary of training forms, all of the models were trained on an input file containing all of the unique word forms with non-zero frequency in the CELEX corpus (Baayen, Piepenbrock, and van Rijn 1993), in phonemic transcription.[11] Since CELEX contains many "duplicate" entries (the same word broken up into two separate entries), these were re-combined into a single type and their token frequencies were summed. Parallel simulations were also carried out with training sets consisting of just lemmas or just monosyllables, but these tended to yield slightly worse results, and are not reported here. Finally, since CELEX uses British transcriptions, it is possible that for purposes of modeling American speakers the training set may lead to inaccurate predictions for certain novel words. As we will see below, this does not appear to be a significant issue for this particular data set (data set 2 below).

3.1 Data set 1: Bailey and Hahn (2001)

The first data set modeled comes from Bailey and Hahn's (2001) study of wordlikeness judgments.[12] It consists of 259 monosyllabic non-words of generally moderate acceptability; examples include *drolf* [drɒlf], *smisp* [smɪsp], *pruntch* [prʌnt͡ʃ], *stulf* [stʌlf], *zinth* [zɪnθ], and *glemp* [glɛmp]. The words were designed to not contain any overt phonotactic violations, though a number of items did contain sequences of dubious legality in English, such as [ʃt#] (*gwesht* [gwɛʃt]) or sNVN (*smimp* [smɪmp]). In the task modeled here, words were presented auditorily in a carrier sentence ("*Zinth*. How typical sounding is *zinth*?"), and participants rated non-words on a scale from 1 (low) to 9 (high). Ratings were rescaled and submitted to an arcsine transformation.

The results are shown in Figures 1–2. We see that none of the models do especially well, although on the whole the lexical models achieve numerically better correlations (Pearson's *R*). Comparing the scatter plots in Figure 1 with those in Figure 2, however, we see that the numerical success of the lexical models may be artificially inflated somewhat by a sparse group of outliers in the upper right.

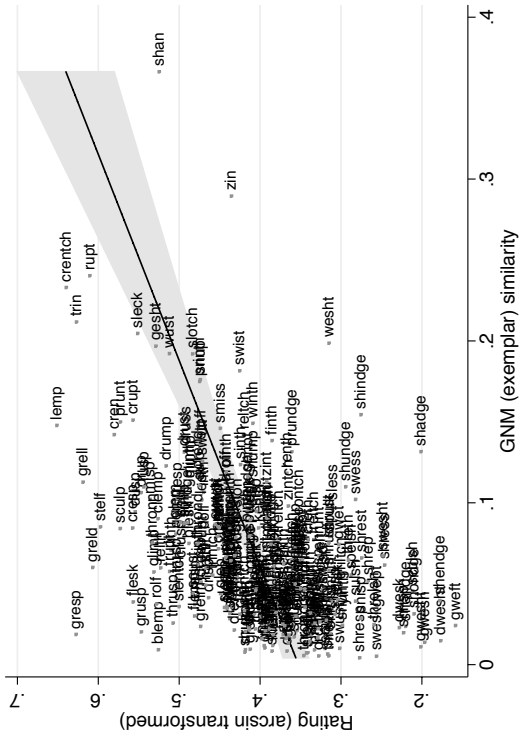
On the whole, it is difficult to conclude much from the graphical or numerical comparisons here, though two points are worth noting. First, this result echoes Bailey and Hahn's own finding

[9] Bailey and Hahn use *average n*-gram transitional probabilities. These were also considered, but yielded uniformly worse results, and are not reported here.

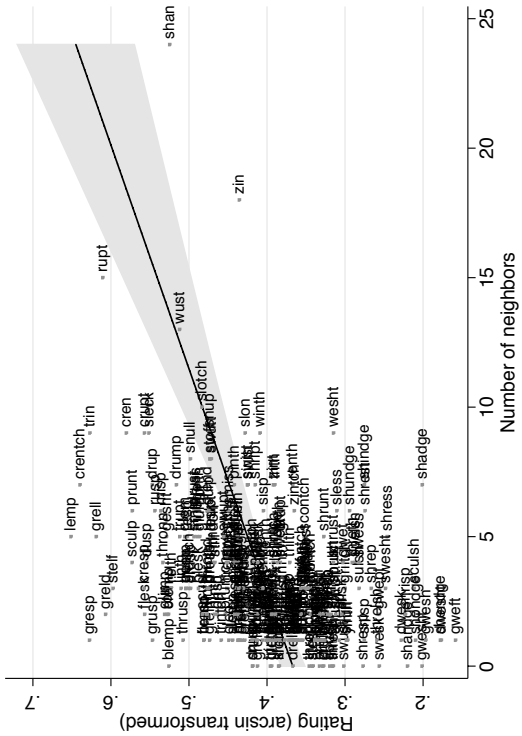
[10] Calculated via web interface: http://www.people.ku.edu/~mvitevit/PhonoProbHome.html

[11] CELEX also contains some word forms with frequency of 0. These were excluded, as they constitute a meaningless and very likely *not* random sample from among the vast set of actual and potential English word forms that do not occur in the British National Corpus. Bailey and Hahn's simulations excluded polysyllabic word forms, for computational reasons; the simulations reported here include polysyllabic forms.

[12] Many thanks to Todd Bailey for making this data available for modeling purposes.

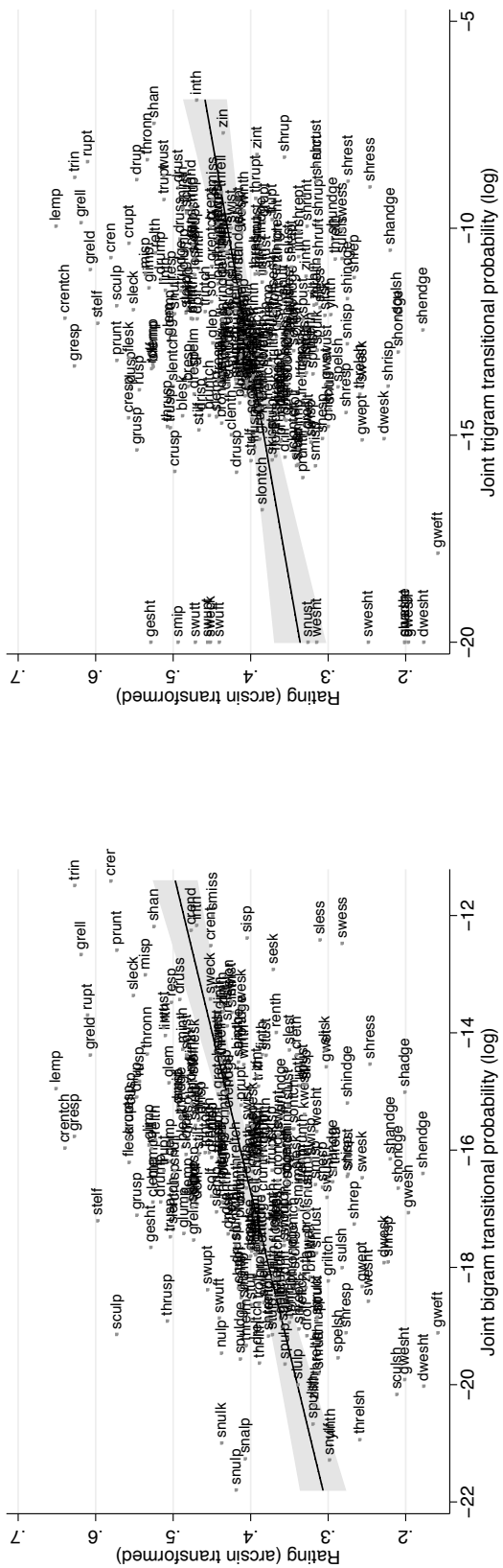


(b) Generalized Neighborhood Model ($r = 0.455$)



(a) Neighborhood density ($r = 0.385$)

Figure 1: Modeling results for Bailey and Hahn (2001) data: lexical models



(a) Joint biphone transitional probability ($r = 0.406$)

(b) Joint triphone transitional probability ($r = 0.262$)

(c) Positional biphone frequency ($r = 0.248$)

(d) Natural class based likelihood ($r = 0.298$)

Figure 2: Modeling results for Bailey and Hahn (2001) data: phonotactic models

that no single model captures their data to any great extent. This result is perhaps not terribly surprising, given that all of the models under consideration are admittedly quite crude and involve many simplifying assumptions. However, there are also reasons to wonder whether Bailey and Hahn's experimental set-up might have introduced additional sources of variability. Their experiment was relatively long and for the most part involved words of intermediate acceptability, with few or no nonce items falling at the endpoints of the scale (canonical word types or items with overt phonotactic violations). It seems possible, therefore, that subjects in their experiment may have had difficulty anchoring their responses to numerical points on the scale in a consistent way, and may have been guided at least partly by local differences between immediately adjacent items.[13] Furthermore, subjects were not asked to repeat or write down what word they had perceived, leaving no way to exclude ratings of the "wrong" (i.e., misperceived) item. Misperceptions could distort the data particularly for words containing non-strident fricatives (e.g., [f] and [θ])—and indeed, examination of the modeling results for the natural class based model shows that for several [θ]-initial words, the model predicts low scores while subjects gave intermediate or high average ratings (e.g., *thrusp* [θrʌsp], *threlm* [θrɛlm], *thrupt* [θrʌpt]). Of course, this may simply be a failing of the model. However, we would be more confident in this conclusion if we knew for certain that subjects had in fact been rating [θ]-initial verbs as intended, and not misperceived [f]-initial versions.

The second thing to note concerns the relative performance of lexical vs. phonotactic models. As Bailey and Hahn also observe, the Generalized Neighborhood model provides the greatest predictive power ($R^2 = 20\%$), followed by biphone transitional probabilities ($R^2 = 16\%$)[14] The predictions of the two types of models also overlap to a large extent ($r = .531$), meaning that once one has been added to the overall model, very little remains for the other to explain (roughly 3% to 6%)[15] Although these results show significant independent contributions of both types of knowledge, it is difficult to reason about their relative importance. Bailey and Hahn argue that the contribution of phonotactic knowledge is quite small, based on the fact that neighborhood similarity is the overall greater effect, and once it is taken into account, phonotactic knowledge explains only a small amount of the remaining variance. There is no a priori reason to interpret the results in this way, however. If we accept the conclusion that both types of knowledge are at play in shaping acceptability judgments and do not impose the requirement that the variance must be explained exclusively by one or the other, then we could argue that neighborhood effects and phonotactic effects are of roughly equal importance (20% vs. 16%, respectively).

Comparing Figures 1 and 2, one also observes a striking difference in the distribution of predicted values. Whereas the phonotactic models in Figure 2 tend to assign scores along a wide range of predicted values (X-axis), the neighborhood models show a significant clustering at the

[13]Bailey and Hahn used four different presentation orders, and report a relatively high effect size across subjects of $\omega^2 = .21$. This value does indeed show that subjects found consistent differences between different non-words, but does not help us distinguish how much of that variance was due to lexical/phonotactic differences, and how much was due to confounding factors.

[14]These values are similar, though not identical to the 22% and 15% that Bailey and Hahn find in their own calculations.

[15]Bailey and Hahn report results from a variety of different combinations of factors. Rather than repeating their full analysis, I focus here on simpler analyses that incorporate only the single best lexical and phonotactic models.

floor of the scale. For reasons discussed by Bailey and Hahn and taken up in more detail below, such skewed distributions pose a challenge for comparing performance across different sets of data. Although the nonwords for this study were designed with care to evoke a normal distribution of subject ratings, not all models make normally distributed predictions. This makes it rather difficult to compare the absolute magnitude of the contribution of different factors using multiple regression.

In sum, the results of this section, while largely replicating those of Bailey and Hahn (2001) in finding significant effects of both lexical and phonotactic knowledge, also raise questions about how to assess the relative importance of these effects. The phonotactic models considered here do not do better than those considered by Bailey and Hahn, though the best one (joint biphone transitional probability) does approximately as well as what they report for average biphone transitional probability. The more sophisticated model that parses strings into natural classes does not do better on this data than a model that phonemes as atomic units, though it does do better than the positional biphone probability metric commonly used in the literature. For reasons just discussed, it is somewhat difficult to interpret the relative performance of the models on this data, and we are also left with the larger question of why ratings in this task were so variable. In the next section, we will see a data set that does not contain so much uninterpretable variance, and therefore may provide a better testing ground for the relative contribution of different types of knowledge in shaping gradient phonological acceptability judgments.

3.2 Data set 2: Albright and Hayes (2003)

The second data set consists of acceptability ratings for a set of 92 non-words collected by Albright and Hayes (2003), in a pre-test for a past tense study. Compared to the Bailey and Hahn test items, the Albright and Hayes words cover a relatively broader range of phonotactic plausibility: 62 of them were estimated ahead of time to be relatively acceptable (containing moderately or very frequent onsets and rhymes—e.g., *kip* [kɪp], *stire* [staɪr], *pank* [pæŋk], *fleep* [fliːp], *blafe* [bleɪf]), while the remaining 30 items were degraded to varying extents by containing unusual or phonotactically marginal sequences (e.g., [pwʌdz] (*pw), [θɹɔɪks] (*[ɔɪk]), [ʃwuʒ] (*ʃw), [skɪk], [snʌm] (*sC_1VC_1, *sNVN)). A few of the 30 less noncanonical items contained completely unattested rhymes (e.g., *smairg* [smɛrg], *smeelth* [smiːlθ][16]) One very un-English item (*bzarshk* [bzarʃk]) was used during training as an example of a word that most English speakers felt could not be a possible word.

In the Albright and Hayes (2003) study, words were presented auditorily in random order in a carrier sentence ("*Blafe*. I like to *blafe*.") Participants repeated the word aloud, and rated it on a scale from 1 ("impossible as an English word") to 7 ("would make a fine English word"). Repetitions were transcribed by two phonetically trained listeners, and if at least one transcriber felt that the subject had repeated the word incorrectly, the rating for that trial was excluded from

[16]These nonwords may be familiar to readers as "distant pseudo-regular" items from Prasada and Pinker's (1993) study of English past tenses.

the analysis. In light of the discussion above regarding potentially noisy data, it is worth noting that participants in this study showed a relatively high degree of between-subject agreement; the correlation between participants 1–10 vs. 11–19 was $r = .864$. The ω^2 value was 0.45, indicating substantial word-by-word differences relative to noise.

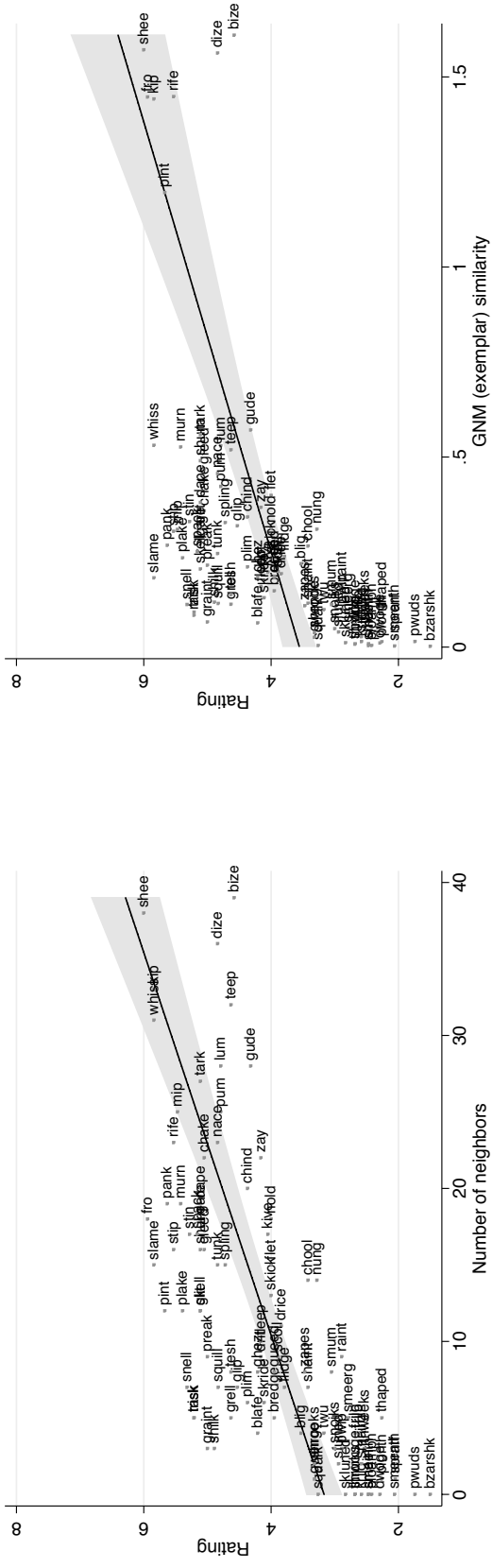
The results are shown in Figures 3–4. Both of the lexical models in Figure 3 perform better on this data than on the previous data set. In both cases, the ratings rise quickly with small amounts of lexical support and then level off with additional support. This suggests that there could be a ceiling effect on subjects' ratings (the fixed 1 to 7 scale preventing higher ratings for better items), or alternatively, that the relation between lexical support and wordlikeness ratings is non-linear. As a result, we must treat the quantitative linear fit (r value) with caution, since it may be artificially lowered by this leveling off effect at the top of the scale.[17] Interestingly, the traditional simplistic definition of neighborhood density outperforms the refined Generalized Neighborhood Model, though the difference is not enormous and may simply reflect the non-linear nature of the effect. More important, both lexical models perform quite poorly at the low end of the scale, failing to distinguish among a the set of unusual words and assigning them all scores at or near zero. The experimental subjects, on the other hand, showed distinct preferences for some of these words over others, as can be seen in the vertical range of ratings for the batch of words at the left of the charts (average ratings ranging from 1.5 to 3.5). This appears to reflect a fundamental inability of lexical models to distinguish adequately between unusual vs. phonotactically ill-formed words—a point that will be discussed in greater detail in section 4 below.

Turning to the phonotactic models in Figure 4, we see that the Vitevitch and Luce (2004) model, which sums over biphone frequencies rather than multiplying to calculate a joint probability, does rather poorly (Figure 4c). Joint biphone transitional probabilities, on the other hand, provide a good fit to subject ratings ($r(89) = .775$), as does the model based on natural classes (Figure 4d: $r(89) = .763$). Both of these models succeed in capturing more than half of the variance in subjects' ratings ($R^2 > .5$). Furthermore, their performance is reasonably consistent across the entire range, requiring no appeals to ceiling or floor effects in the data.

In order to test for independent lexical vs. phonotactic effects, a multiple regression was performed. Since the joint biphone transitional probabilities were the best predictor from individual regressions, they were entered first at $R^2 = 60\%$. GNM similarity scores contributed an additional R^2 of 3.4%, which was not a significant gain ($F(1,89) = 3.14$, $p = .08$).[18] As discussed above, there is no a priori reason to believe that lexical or phonotactic knowledge has "prior" explanatory status, so it is worth considering whether the same result could be obtained with neighborhood similarity alone. When the model was run in the opposite direction, we start with GNM similarity ($R^2 = 33\%$), but in this case biphone transitional probabilities are still able to contribute significant explanatory value ($R^2 = 28\%$, $F(1,89) = 35.38$). Thus, we find that neighborhood similarity

[17] It would not be difficult to get a better estimate of the fit to the data by adopting non-linear models. This option was not explored because both models suffer from a more serious problem at the low end of the scale, discussed immediately below, which could not be remedied with non-linear regression.

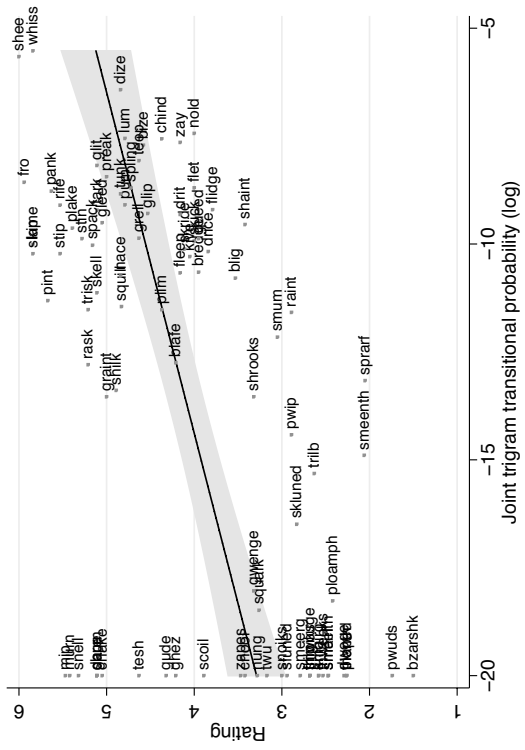
[18] An almost identical result was obtained when traditional neighborhood density counts were entered instead of GNM similarity values.



(a) Neighborhood density ($r = 0.675$)

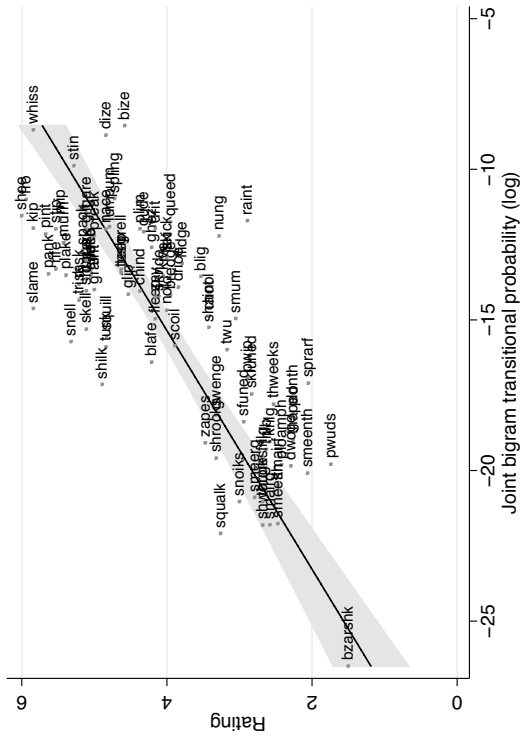
(b) Generalized Neighborhood Model ($r = 0.560$)

Figure 3: Modeling results for Albright and Hayes (2003) data: lexical models



(a) Joint biphone transitional probability ($r = 0.775$)

(b) Joint triphone transitional probability ($r = 0.553$)



(c) Summed positional biphone frequency ($r = 0.204$)

(d) Natural class based likelihood ($r = 0.763$)

Figure 4: Modeling results for Albright and Hayes (2003) data: phonotactic models

is dispensable as a source of explanation for this set of ratings, but phonotactic knowledge is not. This result is almost the opposite of Bailey and Hahn's result, in which neighborhood models were the most predictive and phonotactic models contributed a smaller, but significant predictive power.

To summarize: unlike the previous section, where lexical models showed slightly better performance, for the Albright and Hayes (2003) data phonotactic models come out significantly ahead and no significant neighborhood effects could be observed. A traditional bigram model performed approximately as well as the model that that generalizes by parsing novel strings into combinations of natural classes, though reasons to believe that this would not generally hold true over larger sets of words are discussed below. In the next section, I will consider some possible reasons for the difference between the results for this data as opposed to Bailey and Hahn's data, as well as some additional considerations that favor a phonotactic model based on natural classes as an account of gradient acceptability judgments.

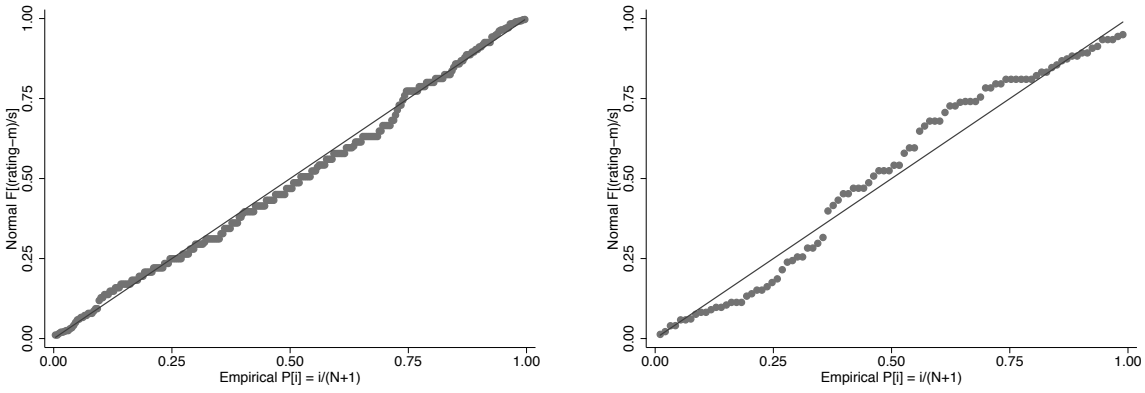
4 Discussion

In the preceding section we saw a puzzling discrepancy: for ratings collected by Bailey and Hahn (2001), lexical models outperformed phonotactic models, though but no model did particularly well. For ratings collected in the Albright and Hayes (2003) study, on the other hand, a phonotactic model that makes reference to the likelihood of various combinations of natural classes is superior. There are several questions to be addressed: why is there an overall difference in performance between the two data sets, and why does there appear to be a qualitative difference between the two studies in whether lexical or phonotactic models work best? What do these results tell us about the mechanisms that give rise to gradient acceptability intuitions? And what kinds of data are needed to go beyond the limitations of these existing data sets?

4.1 Overall variability

The first issue concerns the overall difference in how amenable the two data sets are to modeling. Bailey and Hahn, in discussing the relatively large proportion of the variance in their data that cannot be accounted for by any model under consideration, hypothesize that the difference between their data and other existing studies may be due primarily to the fact that their stimuli were selected to embody a random distribution of acceptability values. They observe that other studies, such as Frisch, Large, and Pisoni (2000), tend to select words that fall at extreme endpoints of the range (canonical or ill-formed), or fall along a flat distribution. They correctly point out that sampling along a flat or dichotomous distribution inflates R^2 values, making it impossible to compare the overall level of fit in different studies. This is certainly an issue in the present case, as well. Whereas the ratings in Bailey and Hahn's study obey a normal distribution, the Albright and Hayes items have a heavy-tailed (closer to flat) distribution. This can be seen in the normal probability

plots in Figure 5, which show that the Bailey and Hahn ratings fall closer to the center line (normal distribution). As a result, we must be cautious in attributing any particular significance to the quantitative differences in model fits across the two studies.



(a) Bailey and Hahn (2001) ratings

(b) Albright and Hayes (2003) ratings

Figure 5: Normal probability plots for ratings across the two studies

Nonetheless, if we visually compare the performance of a single model across the two studies, for example the natural class based model (Figure 2 vs. Figure 4d), it does not appear to be the case that the improved performance on the Albright and Hayes data is due solely to the fact that center of the range is undersampled. Words that are predicted to fall into the intermediate range are, for the most part, given intermediate ratings—unlike Figure 2, in which the center of the plot shows ratings along the entire vertical dimension. As mentioned above, one factor that may have made Bailey and Hahn's data more variable is the fact that most of their non-words were of intermediate acceptability, ranging from *gweft* [gwɛft] on the low end to *crendge* [krɛnd͡ʒ] on the high end. This may have had the effect of making the task more difficult to perform, and could have yielded greater overall variability in the responses. Some support for this comes from the fact that word-by-word differences yielded a much larger effect size between subjects in the Albright and Hayes data ($\omega^2 = .46$) than what Bailey and Hahn report ($\omega^2 = .21$). In addition, the fact that low-probability sequences may be repaired perceptually may have systematically increased the ratings for non-words with easily confused segments (e.g., [f] and [θ]). It is difficult to assess the magnitude of such effects, but they provide plausible, if speculative, reasons why data from the Bailey and Hahn study may be overall more resistant to interpretation with a single simple model of wordlikeness or phonotactic probability. The Albright and Hayes data included clearly acceptable and unacceptable non-word endpoints, and eliminated (as much as possible) responses to misheard items. These features of the experimental design may have contributed to a data set that more clearly shows the unconfounded effect of wordlikeness/phonotactic probability.

4.2 Relative importance of lexical vs. phonotactic knowldge

A more important issue concerns the difference between what appears to be a primarily lexical effect in the Bailey and Hahn data, vs. a primarily phonotactic effect in the Albright and Hayes data. Shademan (2006) argues that this is a task-dependent difference, caused by the fact that Bailey and Hahn included real word fillers among their test items. Shademan hypothesizes that the inclusion of real words engages a mode of processing that involves the lexicon to a greater extent than one would find in a task involving exclusively non-words. In support of this idea, Shademan provides experimental data showing that the relative influence of lexical neighbors can indeed be modulated by whether or not real word items are included in the task. If this is right, then the Albright and Hayes data (which did not include real word fillers[19]) would be a more revealing test of whether a probabilistic phonotactic grammar is needed to explain gradient acceptability judgments. It should be noted, however, that when we compare the predictive power of lexical models across the two studies (Figure 1 vs. Figure 3), we do not actually see a decrease when no real words are involved—in fact, quite the opposite. Instead, the comparison reveals that the models show a similar failing in all cases: failure to adequately differentiate among low-probability items, resulting in a clustering of predicted values at the low end of the range (left of the plots).

The inability of lexical models to discriminate between somewhat unusual and completely phonotactically illegal words reveals a deeper reason why they may be fundamentally ill-suited to the task of modeling gradient acceptability. The attempt to distinguish between possible and completely impossible sequences is at the heart of traditional phonological analysis. If an exemplar model did very well at explaining why *pank* [pæŋk] sounds better than *shresp* [ʃrɛsp], but could not explain why *dlap* [dlæp] or *mrut* [mruːt] are worse than either of them, it would hardly constitute a general-purpose explanation for gradient acceptability. This is significant, because similarity-based models are not designed for capturing intuitions about unattested sequences. In fact, novel words may contain phonotactic violations and yet be very similar to a number of existing words. For example, the non-words *frew* [fruː] and *srew* [sruː] both have very many neighbors (*brew*, *crew*, *drew*, *grew*, *roux*, *screw*, *shrew*, etc.), and indeed, the GNM assigns them very similar scores: *frew* = 1.96, *srew* = 1.68 (in arbitrary units). This means that *srew* is predicted to be in the same range as other moderately well-supported words, such as *lum* [lʌm] (1.59), *wiss* [wɪs] (1.59), and *tark* [tark] (1.68), all of which received ratings greater than 5 out of 7 in the Albright and Hayes study. This ignores that fact that *srew* contains the phonotactically illegal sequence **#sr*, found only in very careful/educated pronunciations of *Sri Lanka*. Although neither of the data sets analyzed provides empirical evidence about the goodness initial *#sr* in particular, I suspect that putting *srew* on a par with words like *lum* and *wiss* highly overestimates its goodness. More generally, the Albright and Hayes non-words did include some items with at least mild phonotactic violations (*twoo* [twuː], *pwuds* [pwʌdz], *smum* [smʌm]), and the fact that the lexical models falter on items from the low end in this study appears to reflect the fact that they are simply unable to encode phonotactic violations (*[+labial]w, *C*w*+round vowel, etc.).

[19]The Albright and Hayes items did include two items that exist as words, but not as verbs: *shee* [ʃiː] and *frow* [froʊ]. In addition, the word *kip* [kɪp] exists as a noun in some dialects, but not the California dialect spoken by Albright and Hayes' subjects. It is not clear whether Shademan's account predicts that the inclusion of just one or two real words should force a task-wide shift towards lexical involvement, or whether it is a matter of degree.

A more concrete demonstration of this can be obtained by focusing on the set of items for which the GNM predicts values that are too high. These were found by fitting the GNM predictions to the subjects' ratings from Albright and Hayes (2003), and calculating the residuals—i.e., the amount by which the model was off in its predictions. The non-words with the greatest positive residuals, or those which the GNM most serious overestimated, are shown in (12).

(12) Non-words for which the GNM most seriously overestimates goodness

throiks [θɹɔɪks], *shwouge* [ʃwuʒ], *rin't* [raɪnt], *frilg* [frɪlg], *krilg* [krɪlg], *smairg* [smɛɹg], *trilb* [trɪlb], *smeelth* [smiːlθ], *smairf* [smɛɹf], *thweeks* [θwiks], *ploamph* [ploʊmf], *dwoge* [dwoʊd͡ʒ], *ploanth* [ploʊnθ], *dize* [daɪz], *thaped* [θeɪpt], *smeenth* [sminθ], *sprarf* [spɹaɹf], *bize* [baɪz], *pwuds* [pwʌdz], *bzarshk* [bzaɹʃk]

The majority of these contain some sort of phonotactic violation, which the GNM is in principle unable to attend to. Moreover, these errors are only the tip of the iceberg. Neither Bailey and Hahn nor Albright and Hayes included significant numbers of words with illegal sequences, so data from these studies allow us to prove only a tiny piece of the overall picture of how well the models account for gradient acceptability. It seems nearly certain that if more illegal sequences had been included, embodying a broader range of more serious phonotactic violations, this inability of neighborhood models to systematically penalize impossible sequences would cause them to suffer. The fact that lexical models do not perform as well as phonotactic models on the Albright and Hayes data may simply be a result of the fact that more mild violations were contained in this set of items, providing a more comprehensive test of the model.

The conclusion, then, is that lexical neighborhood effects are insufficient to explain the gradient differences in acceptability that subjects express in rating non-words. A probabilistic grammar of co-occurrence restrictions is also needed, and indeed, may play the major role in explaining speaker intuitions of wordlikeness. This directly contradicts the conclusion that Bailey and Hahn draw from their own data, but we have seen that this may be a result of focusing on just one portion of the task (differentiating among mostly attested sequences), testing too simple a model of phonotactics (average bigram/trigram transitional probabilities), and perhaps also the inclusion of real words among non-words in the task.

It is also instructive to examine the non-words which the GNM most seriously underestimated, listed in (13). For the most part, these are words that contain legal sequences, but happen to be a bit isolated in the lexicon. Just as speakers are evidently able to focus on phonotactic violations in spite of similarity to existing words in the case of words with illegal sequences, it appears that speakers are also not overly bothered by a lack of many similar existing words in the case of legal sequences.

(13) Non-words for which the GNM most seriously underestimates goodness

slame [sleɪm], *stire* [staɪɹ], *pank* [pæŋk], *snell* [snɛl], *rask* [ræsk], *trisk* [trɪsk], *stip* [stɪp], *plake* [pleɪk], *mip* [mɪp], *wiss* [wɪs], *grin't* [graɪnt], *skell* [skɛl], *spack* [spæk], *stin* [stɪn], *shilk* [ʃɪlk], *squill* [skwɪl], *gare* [gɛɹ], *preek* [priːk], *glit* [glɪt], *murn* [mr̩n]

If this is right, it constitutes additional evidence in favor of a probabilistic grammar distinct from the lexicon, since similarity-based neighborhood models are apparently not sufficient to distinguish even among attested sequences. This batch of words should be uniformly grammatical under a categorical account of English phonology, yet the best account of gradient differences among them is stated not in terms of neighborhood size, but rather, in terms of sequences of phonological entities. In order to maintain the view that "grammar is categorical, performance is gradient", we would thus need to admit three kinds of knowledge: (1) categorical grammar, (2) stochastic phonotactics, stated using the same representational language but with probabilities attached, and (3) lexical knowledge. The simpler theory is one in which (1) and (2) are combined into a single probabilistic grammar that directly regulates the gradient acceptability of sequences. As stated at the outset, this conclusion is by no means a new one—but it is rarely argued for by showing that a purely lexical similarity-bsaed account of the same data fails.

4.3 Are biphones enough?

A notable feature of the results in section 3 is that a model based on counts of phonemes as atomistic units ('*p*', '*æ*', etc.) always outperforms the more sophisticated natural class model proposed in section 2.2. One might be tempted to consider this a negative result: phonological abstractions like features and natural classes, at least in the way they are used here, are useless in modeling phonological acceptability. There is reason to believe that this conclusion is premature, however. Comparing the results of the biphone and triphone models in Figure 4, we see that the triphone model incorrectly predicts values at the floor for very many test items. The reason, as discussed in section 2.2, is that the number of possible trigrams is enormous and many logically possible triphones are, for accidental or for principled reasons, unattested among existing English words. The triphone model is unable to distinguish between accidental gaps (such as [gɛz]) and what would traditionally be termed ungrammatical sequences (such as [bza]). Even within the class of accidentally unattested sequences, subjects are able to discrimate between relatively better ones (e.g., [gɛz]) and worse ones (e.g., [nʌŋ]). The problem of distinguishing among equally unattested items becomes even more pressing if we turn to 'somewhat' vs. 'very' ungrammatical sequences: [bza] vs. [mgl]. Since the data sets analyzed here contained few rare or unattested bigrams, the bigram model actually does quite well on these tasks. This should not be taken as victory for such models, however; one would need only to test for differences between unattested [bw], [dl], [bn], and [bd] to find cases that the model is unable to distinguish. Moreton (2002) shows that speakers do in fact show a preference for [bw] over [dl], in spite of the fact that they are equally unattested.[20]

[20] At least for some speakers of English; loan words like *bwana* or *bueno* have introduced [#bw] sequences, though loans like *Tlingit* have also introduced [#dl]-like sequences.

It is likely that an ability to refer to natural classes is crucial in modeling the way that speakers generalize to previously unseen biphones.

4.4 Token frequency

There is one last respect in which ratings from these two studies appear to bear the hallmarks of a grammatical, rather than lexical effect: none of the models explored here derived any advantage from their ability to weight the contribution of individual data according to their token frequency. This is seen in several ways: the GNM and simpler neighborhood density models performed best when their frequency weighting was turned off; the natural class-based model did best when instantiation costs were calculated without reference to the relative frequency of different segments; and the Vitevitch and Luce model, which intrinsically takes token frequency into account, did not come out ahead by virtue of having this ability. This finding contradicts that of Bailey and Hahn (2001), who found a significant contribution for frequency weighting. The effect they found was very small, however (a 1% gain in performance for this data set), and I have been unable to replicate it. In fact, in most cases, taking token frequency into account makes very little difference in the predictions of the models. The reason is that most words in the lexicon are very low frequency, so a boost for high token frequency words gives more influence to just a small set of items. When it does make a difference, though, it tends to be a deleterious one. This confirms a number of previous claims in the literature that pattern strength is determined by type, not token frequency (Bybee 1995; Albright 2002b; Albright and Hayes 2003; Hay, Pierrehumbert, and Beckman 2004). I take this finding to be at odds with the idea that gradient acceptability arises as a by-product of consulting the lexicon, since lexical access is known to be highly sensitive to frequency, and yet gradient acceptability appears to be completely impervious to it.

5 Conclusion

Based on analysis of currently available data, it appears that gradient phonotactic acceptability bears the markings of a grammatical effect: it requires reference to statements about combinations of sequences stated in terms of features and natural classes, often held to be the representational language of grammars. At the same time, it is insensitive to token frequency, which is widely observed to influence performance. It must be acknowledged, however, that for reasons discussed above, the current data is inadequate to provide a complete test of the models under comparison. Most important, the data sets do not contain sufficient information to calibrate the full range of acceptability, including canonical and illegal sequences. Furthermore, neither study included items designed to test specifically for the contribution of token frequency effects, so the conclusions here must be based tentatively on *post hoc* comparisons. Finally, the items in these studies involve a rather limited rather of word shapes, preventing us from testing the broader range of structures thought to play a role in phonological grammars. The analyses carried out here highlight some of

the ways in which it is important to expand the range of non-words, and control the experimental conditions under which acceptability judgments are elicited. Experimental studies carrying out these goals, and comparing gradient acceptability judgments to other forms of linguistic behavior, are left for future research.

6 Appendix: list of nonce words from Albright & Hayes (2003)

Word	Rating	Word	Rating	Word	Rating	Word	Rating	Word	Rating	Word	Rating
ʃiː	6.00	spæk	5.16	lʌm	4.79	skraɪd	4.11	nʌŋ	3.28	θwiːks	2.53
froʊ	5.94	gɛɹ	5.11	pʌm	4.79	kɪv	4.05	skwalk	3.26	smiːlθ	2.47
kɪp	5.84	ʃɻn	5.11	splɪŋ	4.72	skɪk	4.00	twuː	3.17	smɛrf	2.47
wɪs	5.84	taɪk	5.11	gɹɛl	4.63	flɛt	4.00	smʌm	3.05	ploʊmf	2.42
sleɪm	5.84	deɪp	5.11	tɛʃ	4.63	noʊld	4.00	snɔɪks	3.00	dwoʊd͡ʒ	2.29
pɪnt	5.67	skɛl	5.11	tiːp	4.63	brɛd͡ʒ	3.95	sfuːnd	2.94	ploʊnθ	2.26
pæŋk	5.63	glɪt	5.11	baɪz	4.58	kwiːd	3.95	pwɪp	2.89	θeɪpt	2.26
raɪf	5.53	t͡ʃeɪk	5.05	glɪp	4.53	skɔɪl	3.89	raɪnt	2.89	smiːnθ	2.06
stɪp	5.53	gliːd	5.05	plɪm	4.37	draɪs	3.84	skluːnd	2.83	spɹaɪf	2.05
mɪp	5.47	pɹiːk	5.00	t͡ʃaɪnd	4.37	flɪd͡ʒ	3.79	smiːɹg	2.79	pwʌdz	1.74
mɻn	5.42	graɪnt	5.00	guːd	4.32	blɪg	3.53	θɹɔɪks	2.68	bzarʃk	1.50
pleɪk	5.39	ʃɪlk	4.89	bleɪf	4.21	zeɪps	3.47	fɹɪlg	2.68		
snɛl	5.32	daɪz	4.84	gɛz	4.21	t͡ʃuːl	3.42	ʃwuːʒ	2.68		
stɪn	5.28	tʌŋk	4.84	zeɪ	4.16	ʃaɪnt	3.42	tɹɪlb	2.63		
trɪsk	5.21	neɪs	4.84	drɪt	4.16	gwɛnd͡ʒ	3.32	smɛrg	2.58		
ɹæsk	5.21	skwɪl	4.83	fliːp	4.16	ʃɹʊks	3.32	kɹɪlg	2.58		

References

Albright, A. (2002a). Islands of reliability for regular morphology: Evidence from Italian. *Language 78*(4), 684–709.

Albright, A. (2002b). The lexical bases of morphological well-formedness. In S. Bendjaballah, W. U. Dressler, O. E. Pfeiffer, and M. Voeikova (Eds.), *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000*, Number 218 in Current Issues in Linguistic Theory, pp. 1–8. Benjamins.

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

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

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

Anshen, F. and M. Aronoff (1988). Producing morphologically complex words. *Linguistics 26*, 641–655.

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

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

Bauer, L. (2001). *Morphological Productivity*. Cambridge University Press.

Berko, J. (1958). The child's acquisition of English morphology. *Word 14*, 150–177.

Berwick, R. C. (1986). Learning from positive-only examples: The subset principle and three case studies. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach: Volume II*, pp. 625–645. Los Altos, CA: Kaufmann.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins Publishing Company.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes 10*(5), 425–255.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. L. and E. Pardo (1981). On Lexical and morphological conditioning of alternations: a nonce-probe experiment with Spanish verbs. *Linguistics 19*, 937–968.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York: Harper and Row.

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

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

Cutler, A. and D. Norris (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance 14*, 113–121.

Eddington, D. (1996). Diphthongization in spanish derivational morphology: An empirical investigation. *Hispanic Linguistics 8*, 1–35.

Ernestus, M. and R. H. Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language 79*, 5–38.

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

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

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

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

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

Hammond, M. (1999). *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. Oxford University Press.

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

Hay, J. and H. Baayen (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences 9*, 342–348.

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

Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater, and W. Zonneveld (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge, UK: Cambridge University Press.

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

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*, pp. 145–165. San Diego: Academic Press.

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

Jusczyk, P. W., P. A. Luce, and J. Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language 33*, 630–645.

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

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

Luce, P. A. and D. B. Pisoni (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing 19*, 1–36.

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

Nakisa, R. C., K. Plunkett, and U. Hahn (1997). A Cross-Linguistic Comparison of Single and Dual-Route Models of Inflectional Morphology. In P. Broeder and J. Murre (Eds.), *Cognitive Models of Language Acquisition*. Cambridge, MA: MIT Press.

Newman, R. S., J. R. Sawusch, and P. A. Luce (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance 23*, 873–889.

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

Nosofsky, R. M. (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology 34*, 393–418.

Ohala, J. and M. Ohala (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In *Experimental Phonology*, pp. 239–252. Orlando, FL: Academic Press.

Pierrehumbert, J. (2002). An unnatural process. In *Labphon 8*.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.

Prasada, S. and S. Pinker (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes 8*, 1–56.

Prince, A. and P. Smolensky (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing.

Prince, A. and B. Tesar (2004). Learning phonotactic distributions. In R. Kager, J. Pater, and W. Zonneveld (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge, UK: Cambridge University Press.

Saul, L. and F. Pereira (1997). Aggregate and mixed-order Markov models for statistical language processing. In C. Cardie and R. Weischedel (Eds.), *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp. 81–89. Somerset, New Jersey: Association for Computational Linguistics.

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

Schütze, C. (2005). Thinking about what we are asking speakers to do. In S. Kepser and M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, pp. 457–484. Mouton de Gruyter.

Shademan, S. (2006). Is phonotactic knowledge grammatical knowledge? In D. Baumer, D. Montero, and M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pp. 371–379. Somerville, MA: Cascadilla Proceedings Project.

Tessier, A.-M. (2006). *Biases and Stages in Phonological Acquisition*. Ph. D. thesis, University of Massachusetts Amherst.

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

Vitevitch, M. and P. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language 40*, 374–408.

Vitevitch, M., P. Luce, J. Charles-Luce, and D. Kemmerer (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech 40*, 47–62.

Vitevitch, M., P. Luce, D. Pisoni, and E. Auer (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language 68*, 306–311.

Vitevitch, M. S. (2002). Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and Speech 45*, 407–434.

Vitevitch, M. S. and P. A. Luce (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science 9*, 325–329.

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

Vitevitch, M. S. and P. A. Luce (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory & Language 52*, 193–204.

Vitz, P. C. and B. S. Winkler (1973). Predicting the judged 'similarity of sound' of English words. *Journal of Verbal Learning and Verbal Behavior 12*, 373–388.

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

Zuraw, K. (2000). *Patterned Exceptions in Phonology*. Ph. D. thesis, UCLA.

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