*Phonology* **26** (2009) 9–41. © Cambridge University Press 2009
doi:10.1017/S0952675709001705 Printed in the United Kingdom

# *Feature-based generalisation as a source of gradient acceptability**

**Adam Albright**
Massachusetts Institute of Technology

Phonological judgements are often gradient: *blick* > ʔ*bwick* > **bnick* > ***bzick*. The mechanisms behind gradient generalisation remain controversial, however. This paper tests the role of phonological features in helping speakers evaluate which novel combinations receive greater lexical support. A model is proposed in which the acceptability of a string is based on the most probable combination of natural classes that it instantiates. The model is tested on its ability to predict acceptability ratings of nonce words, and its predictions are compared against those of models that lack features or economise on feature specifications. The proposed model achieves the best balance of performance on attested and unattested sequences, and is a significant predictor of acceptability even after the other models are factored out. The feature-based model's predictions do not completely subsume those of simpler models, however. This may indicate multiple levels of evaluation, involving segment-based phonotactic probability and feature-based gradient phonological grammaticality.

## 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.

* I am very grateful to the following people for valuable comments, suggestions and technical advice: Arto Anttila, Todd Bailey, Andries Coetzee, Edward Flemming, Bruce Hayes, René Kager, Jonah Katz, Joe Pater, Donca Steriade, Josh Tenenbaum, Colin Wilson and two anonymous reviewers. All remaining errors are, of course, my own.

[1] I use the term gradient ACCEPTABILITY rather than gradient WORDLIKENESS or GRAMMATICALITY, since those terms presuppose particular underlying mechanisms.

(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 | [mɹak] |
| *worst* | bzack | [bzæk] | shöb | [ʃœb] |

Numerous studies have demonstrated a relation between gradient acceptability and lexical statistics, using data from different domains and different experimental methodologies (Greenberg & Jenkins 1964, Scholes 1966, Ohala & Ohala 1986, Coleman & Pierrehumbert 1997, Frisch *et al.* 2000, Bailey & Hahn 2001, Hay *et al.* 2004, Shademan 2007 and many others).

The fact that acceptability correlates with facts about the English lexicon is intuitive and unsurprising. For example, counts of English lemmas from CELEX (Baayen *et al.* 1993) reveal that English has fewer words starting with [bl] than with [st], fewer yet starting with [bw] and none starting with [dl] or [bz], roughly mirroring the intuitions in (1). What remains far from certain, however, is what form that knowledge takes, how it is learned and whether gradient differences are reflected in the grammar itself. On the one hand, reactions to nonce words (or 'wug' words; Berko 1958) are often taken to be a powerful source of evidence for the grammatical knowledge that speakers have about licit structures in their language. At the same time, acceptability judgments may also be influenced by a number of other, non-grammatical considerations. First, speakers may estimate the overall phonetic similarity of nonce words to known words, calculating the degree of support from existing lexical items (NEIGHBOURHOOD DENSITY; Greenberg & Jenkins 1964, Coltheart *et al.* 1977, Luce 1986, Ohala & Ohala 1986, Newman *et al.* 1997, Luce & Pisoni 1998, Bailey & Hahn 2001). In addition, speakers may decompose novel words into substrings in the course of word segmentation and lexical access, calculating the probability of each combination of phonemes in the lexicon (PHONOTACTIC PROBABILITY; Jusczyk *et al.* 1994, Vitevitch *et al.* 1997, Vitevitch & Luce 1998, Auer & Luce 2005). Even relatively simple versions of neighbourhood density and phonotactic probability, which employ none of the familiar features of phonological grammar such as prosodic structure, features or markedness, have been shown to be significant predictors of the acceptability of nonce words composed of rare *vs.* common sequences (Ohala & Ohala 1986, Vitevitch *et al.* 1996, Bailey & Hahn 2001). Such results are sometimes taken to indicate that responses to nonce words have little to do with grammar at all, but rather reflect the involvement of non-grammatical mechanisms such as a prelexical parser that calculates phonotactic probability for the purpose of finding word and morpheme boundaries (Vitevitch & Luce 1999, Hay 2003), or a

word-learning mechanism that evaluates the likelihood that a hypothetical new word is actually a word of the language (Schütze 2005).

Although simple neighbourhood and phonotactic models have been moderately successful in predicting gradient acceptability of monosyllables composed of attested phoneme combinations (*stin* ≻ *blafe*), they are not so well suited to the task of modelling differences between unattested combinations (*dlack* ≻ *bzack*). For the task of differentiating among unattested sequences, it appears that models may benefit from incorporating elements of phonological structure. Coleman & Pierrehumbert (1997) propose a model that parses words into maximally probable syllables, allowing unattested medial clusters to be broken up into the most probable coda + onset combination. This strategy has been shown to outperform simpler measures like neighbourhood density or biphone probability in modelling acceptability of medial sequences in polysyllabic words (Coleman & Pierrehumbert 1997, Frisch *et al.* 2000, Hay *et al.* 2004). In order to capture differences among unattested initial and final clusters ([#dl] ≻ [#bz]), for which the syllabic parse is unambiguously the same, it is useful to refer to a different type of phonological structure, namely, featural overlap with attested clusters such as [#bl], [#gl] and [#dɹ]. Hayes & Wilson (2008) propose an inductive model that uses features to learn constraints on natural classes of phonemes, which allows different unattested clusters to be assigned different penalties. They show that this model outperforms simple neighbourhood density, bigram transitional probability and the Coleman & Pierrehumbert (1997) model in predicting the acceptability of a variety of novel onset clusters. This success on unattested clusters is a direct consequence of the model's use of phonological features. However, as we will see below, the model fails to differentiate clusters that are at least moderately well attested ([#st] ≻ [#bl]).

In this paper, I propose a model that combines the advantages of a simple segment-based *n*-gram phonotactic model in capturing differences among attested sequences with the advantages of a feature-based model in generalising to unattested sequences. Like an *n*-gram model, the model decomposes words into strings of a fixed length (here biphones) and calculates the transitional probability from one segment to the next. However, rather than considering transitions between segments directly, the model calculates probabilities over combinations of natural classes. For each biphone in the word, the model determines the most probable parse in terms of natural classes, assigning a score based on the most probable combinations. Following Hayes & Wilson (2008), the performance of the model is first assessed on a set of acceptability ratings of onset clusters (Scholes 1966). The results (§3.1) show that the model provides a reasonably good balance of performance on gradient acceptability of both attested and unattested clusters. In §3.2, the model is tested on its ability to model acceptability ratings of longer strings, using data from a set of monosyllabic nonce words involving a mix of common and rare sequences (Albright & Hayes 2003). These results show that ratings of entire words

are captured quite accurately using a model that combines probabilities cumulatively across the word, echoing findings by Coleman & Pierrehumbert (1997), Frisch *et al*. (2000) and others, but contrary to standard assumptions of Optimality Theory (Prince & Smolensky 2004). Importantly, this section also finds distinct effects of segment-based and feature-based biphone probability, which apparently cannot be reduced to one another. §4 reconciles the current findings with those of similar recent studies, and discusses the role that such models play in motivating probabilistic phonotactic grammar.

## 2 A feature-based model of biphone probability

Following standard terminology in the psycholinguistics literature, we consider a model to be a PHONOTACTIC MODEL if it evaluates the acceptability of a word by decomposing it into substrings of (possibly non-adjacent) phonological material, and then combining the assessments of subparts to yield a prediction for the entire word. Phonological grammars, as embodied for example in the rule-based approach of *SPE* (Chomsky & Halle 1968) or the constraint-based approach of Optimality Theory (Prince & Smolensky 2004), represent sophisticated hypotheses about the types of knowledge about substrings that may be encoded and about how acceptability is computed. However, before we consider the contribution of phonological structure to a phonotactic model, it is useful to consider first as a baseline a simple and widely used strategy for modelling words as series of independent subparts: *n*-gram models, which keep track of the probabilities of substrings of *n* elements (see Jurafsky & Martin 2000: ch. 6 for an overview).

Numerous variants of *n*-gram models have been used to predict the phonotactic acceptability or probability of words. In principle, substrings of any length *n* could be considered, but in practice, it does not appear to be practical or useful in phonology to consider substrings longer than two or three segments (biphones, triphones), since the number of logically possible longer strings is enormous, and the probability of encountering any one of them more than a few times in the lexicon becomes very small. For example, the novel word *dresp* [dɹɛsp] contains the tetraphone [ɛsp#], which happens to be unattested in English.[2] This gap appears to be treated as a statistical accident by native speakers, however, and nonce words such as [dɹɛsp] are quite acceptable. More generally, for longer sequences, the number of possible *n*-grams grows exponentially, meaning a much larger corpus is needed to estimate their frequencies with any degree of certainty. In the present case the corpus is the lexicon of attested words, which is of limited size. Hence, in the usual case it is not possible to get more data

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

| biphone | stɪn | | bleɪf | |
|---|---|---|---|---|
| 1 | #s | 0·118 | #b | 0·057 |
| 2 | st | 0·205 | bl | 0·106 |
| 3 | tɪ | 0·192 | leɪ | 0·042 |
| 4 | ɪn | 0·108 | eɪf | 0·007 |
| 5 | n# | 0·151 | f# | 0·067 |
| log (prob) | | –4·12 | | –6·93 |

*Table I*
Biphone transitional properties of *stin* and *blafe*.

about the goodness of a particular *n*-gram by collecting more words.[3] Bailey & Hahn (2001) find that even triphones are of little or no use in predicting acceptability of English monosyllables, though taking at least triphones into consideration would probably be advantageous in evaluating polysyllabic words.

Given a particular length of substring (*n*), there are also various ways to calculate the probability of each *n*-gram; we focus here on the TRANSITIONAL PROBABILITY from the first $(n-1)$ segments to the final segment. For example, given a word *abcd* with biphones #*a*, *ab*, *bc*, *cd*, and *d*#, the transitional probability of *ab* (that is, the conditional probability of *b* given *a*) is defined as the number of times *ab* occurs in the corpus divided by the total number of times *a* appears as the first member of a biphone (i.e. the proportion of occurrence of *a* that are followed by *b*).[4]

Biphone probabilities in one form or another have been shown to correlate with a wide variety of phenomena, including phoneme identification (Pitt & McQueen 1998), repetition speed in shadowing tasks (Vitevitch *et al*. 1997, Vitevitch & Luce 1998, 2005), response time in same/different tasks (Vitevitch & Luce 1999) and looking times in 9-month-olds (Jusczyk *et al*. 1994). Furthermore, even a simple model employing just biphone transitional probabilities can make a certain amount of headway in capturing gradient differences in nonce word

[3] The fact that words tend to be short further limits the usefulness of longer substrings. The median length of single-word lemma entries in the CELEX corpus (Baayen *et al*. 1993) is six segments, meaning that on average, each word contributes information about just six trigrams (including word edges).

[4] Other strategies have been proposed, including simply summing over the component biphone probabilities Vitevitch & Luce (1998, 2004), but joint probabilities (products) yield better results for all of the data sets discussed below. It is also possible to normalise for length (in segments) by taking the arithmetic mean or geometric mean (Bailey & Hahn 2001) of the probabilities. However, Bailey & Hahn (2001 : 582) report that averaging actually decreased the performance of their model slightly, and similar results were found for the data to be discussed in §3.2.

| biphone | bnæk | | bdæk | | bzæk | |
|---|---|---|---|---|---|---|
| 1 | #b | 0·057 | #b | 0·057 | #b | 0·057 |
| 2 | bn | 0·003 | bd | 0·004 | bz | 0·004 |
| 3 | næ | 0·015 | dæ | 0·017 | zæ | 0·006 |
| 4 | æk | 0·104 | æk | 0·104 | æk | 0·104 |
| 5 | k# | 0·147 | k# | 0·147 | k# | 0·147 |
| log (prob) | | –7·43 | | –7·19 | | –7·68 |

*Table II*
Biphone transitional properties of *bnack*, *bdack* and *bzack*.

acceptability (Vitevitch *et al*. 1997, Bailey & Hahn 2001). For example, the difference between *stin* and *blafe* emerges straightforwardly as a difference in the transitional probability of their biphones. The higher probability of *stin* is illustrated in Table I, using biphone probabilities calculated over all lemma entries with non-zero token frequency in CELEX.

In spite of such successes, the limitations of simple biphone models make them inadequate for modelling many aspects of phonological knowledge, including the preferences that speakers show for some unattested clusters over others. For example, Berent *et al*. (2007) document a preference among English speakers for onset clusters with rising sonority: [#bn] ≻ [#bd], [#bz]. As shown in Table II, a simple bigram model does not predict this preference, since all three [bC] sequences have comparably low transitional probability (based on word-medial occurrences).

Unattested onset clusters pose two distinct problems for the simple biphone model. The first is that a model has no way to distinguish onset clusters from intervocalic (heterosyllabic) clusters, with the consequence that unattested onset clusters are assigned probability according to their medial occurrences. It would be possible to distinguish initial from word-medial clusters by considering triphones ([bn] *vs*. [bd]) or by incorporating syllable structure ([$_{\sigma}$bn *vs*. [$_{\sigma}$bd) or position in the word (initial, second, third; Vitevitch & Luce 2004), but these strategies raise a second problem: as word-initial/onset clusters, all three have a frequency of 0 in CELEX. Some general strategies for dealing with completely unattested sequences include reserving some probability mass for unseen sequences (smoothing), or reasoning about unseen sequences based on the probabilities of their shorter subparts (back-off); see Jurafsky & Martin (2000: ch. 6) for an overview. In this paper, we test a different approach, using phonological structure to generalise to unattested sequences: we attempt to derive the preference for [#bn] ≻ [#bd] as a consequence of the large number of phonological features that [#bn] shares with existing clusters such as [#bl] or [#sn], and the lack of existing clusters that are directly

comparable to [#bd]. Specifically, we propose a model that calculates the probability not just of individual biphones ([bn], [bd]), but also of combinations of natural classes ([−sonorant][+nasal], [−continuant][+sonorant], etc.). This allows the model to generalise to unseen combinations by assigning them phonological structure, parsing them as likely or unlikely combinations of natural classes.

The insight of generalisation 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: 115). Similarly, words beginning with [#bn] are unattested in English, but words beginning with [#bl] and [#bɹ] (i.e. initial [b] preceding other coronal sonorants) do occur. This may lead speakers to conclude that words beginning with [#bn] are at least faintly conceivable in English. The task of the model, then, is to determine what features the attested combinations have in common, and estimate the likelihood that sequences involving members of a broader class (front lax vowels, coronal sonorants) could also occur. As these examples show, not all generalisations are equally likely. In particular, comparison of sequences like [ɪsp] and [ɛst] strongly supports novel [ɛsp], while comparison of sequences like [#bl] and [#sn] provides only slight support for [#bn].

One proposal for how speakers compare sequences to discover more general patterns of natural classes is the MINIMAL GENERALISATION approach (Albright & 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, so that the algorithm conservatively avoids generalisation to unseen feature values. For example, comparison of [ɪsp] and [æsp] yields a generalisation covering all front lax vowels (2a), while further comparison with [ɛsk] extends the generalisation to all front lax vowel + *s* + voiceless stop sequences (2b).[5]

[5] 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 V*st* sequences. In general, bivalent (equipollent) feature definitions limit abstraction, by creating more 'negatively specified' natural classes such as [−coronal] and preventing generalisation to unseen positive values. For present purposes I will assume privative place features.

(2) *Minimal generalisation*

| | | | | |
|---|---|---|---|---|
| a. | | ɪ | s | p |
| | + | æ | s | p |
| | → | [−back, −round, −tense] | s | p |
| b. | + | ɛ | s | k |
| | → | [−back, −round, −tense] | s | [−son, −cont, −voice] |

In the example in (2), the sequences under comparison differ rather minimally in their feature values, and the resulting generalisations 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 (3). 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.

(3) *A minimal but sweeping generalisation*

| | | æ | s | p |
|---|---|---|---|---|
| | + | b | oʊ | ə |
| | → | [+voice] | *seg* | *seg* |

The pattern in (3) is well attested in English, since voiced segments are extremely 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* [bzaɹʃk] should be very acceptable, since the sequences [bza], [aɹʃ] and [ɹʃk] get a good deal of support from attested '[+voice] *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 [boʊə] provide little or no support for [bza] or [ɹʃk]. 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 due 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 a large number of 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 one-third. 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 '[+voice] *seg seg*' sequence, then the chance of getting attested [ɪsp] or [æsp] at random from among all logically possible '[+voice] *seg seg*' combinations is tiny. Although both patterns can describe the existing data, the more specific characterisation 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 characterise 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.[6] It is also related to proposals within Optimality Theory for seeking the most restrictive possible grammar (Hayes 2004, Prince & Tesar 2004, Tessier 2006) as a way of obeying the subset principle (Berwick 1986).

In order to test the idea that generalisation by natural classes requires a balance of frequency and specificity, 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 (4a), then the probability of the segments *ab* as members of natural classes *A*, *B* respectively can be defined as in (4b).[7]

[6] For an MLE-based approach to *n*-gram models that refer to classes of items, see Saul & 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 characterised in different terms with respect to what occurs before it *vs.* after it. Ultimately, we seek a learning algorithm that incorporates the principles of MLE, without the assumptions about class structure that existing class-based *n*-gram models typically make.

[7] Since natural classes in phonology are characteristically overlapping and each segment belongs to many natural classes, the equation in (4a) will not yield values that sum to 1 over all segments and natural classes; a normalisation term would be needed to convert these scores to true probabilities.

(4)

a. Probability of the sequence $ab = \dfrac{\text{Number of times } ab \text{ occurs in corpus}}{\text{Total number of biphones}}$

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

$$\propto \frac{\text{No of times } AB \text{ occurs in corpus}}{\text{Total number of biphones}} \times \text{P(choosing } a|A) \times \text{P(choosing } b|B)$$

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 (5a). 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 (5b). For the simulations reported below I assume the simpler definition, which is not sensitive to token frequency.

(5) *Two definitions of instantiation costs*

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

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

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 '[+voice] *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 (5a)), 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 $\approx 1{\cdot}48 \times 10^{-5}$, and a correspondingly low predicted probability. We 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 & Pierrehumbert 1997 and Albright & Hayes 2002).

An important feature of this way of selecting the right level of generality at which to parse a given sequence into natural classes is that the model must maximise not only the probability of the combination of classes, but also instantiation probability. This leads the model to favour parses involving natural classes that are as specific as possible, since the probability of instantiating a specific natural class as one of its (few) segments is much higher than the probability of randomly selecting the same segments from a very large natural class. This has the effect that for attested sequences, the model tends to prefer the most specific parse possible – namely, the *n*-gram referring to the precise combination of segments that is being evaluated. For example, in parsing the biphone [pɹ], the model selects the natural class combination [−continuant, +labial, −voice][+rhotic] – i.e. [p][ɹ]. Parses in terms of broader natural classes come at a cost, and are favoured when the probability of the specific biphone is very low. As we will see, this allows the model to combine some of the virtues of *n*-gram models for modelling attested sequences, while making use of feature-based generalisation for unattested sequences.

To summarise, 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 (4b). Taken together, the need to maximise high featural *n*-gram probabilities and minimise instantiation costs should drive the model to seek characterisations 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 model

### 3.1 Onset-cluster acceptability

Following Hayes & Wilson (2008), we first test the model on its ability to predict English speakers' acceptability ratings of existing and novel onset clusters, as documented by Scholes (1966: Experiment 5). In Scholes' study, nonce monosyllables involving a mix of attested and unattested onset clusters paired with fairly common rhymes were played to 33 seventh-graders, who were asked to indicate whether the word was 'usable' as a word of English (binary forced choice). The results show that attested clusters are (unsurprisingly) almost always deemed better than unattested clusters.[8] However, there are also considerable differences among both the attested and unattested clusters, illustrated with representative examples in (6).

[8] The sole exception in the data is [#sk] (in the word *skeep* [skiːp]), which was deemed acceptable by fewer subjects than [#mɹ] (*mrun* [mɹʌn]) or [#ʃl] (*shlurk* [ʃlɻk]) were.

(6) *Cline of acceptability for attested and unattested onset clusters*

| | | | | | |
|---|---|---|---|---|---|
| attested: | kɹʌn, stɪn | 100% | unattested: | nlʌŋ, sɹʌn | 47% |
| | blʌŋ, slɨk | 84% | | nɹʌŋ, zɹʌn | 33% |
| | glʌŋ, ʃɹʌn | 72% | | vtɪn, fnɛt | 19% |
| | | | | ʒpeɪl, vmæt | 11% |
| | | | | vkiːp, ʒviːl | 0% |

Hayes & Wilson (2008) model these differences as a function of the onset clusters alone, ignoring any potential contribution of the remainder of the word (e.g. *__eep vs. __urk*). They trained a variety of models on a corpus of word-initial onsets, adapted from the CMU pronouncing dictionary[9] by removing onsets such as [#zw] and [#sf] that occur primarily in borrowed words. Comparing correlations between the models' predictions and the proportion of 'yes' responses for each cluster, they find that a simple (segment-based) *n*-gram model does not achieve as good a fit as their inductive feature-based model, which uses natural classes to distinguish among unattested clusters. This provides strong *a priori* support for the idea that features are useful in modelling how speakers generalise from attested to unattested sequences. In this section, we test how the feature-based model proposed in the previous section compares in its ability to model the observed differences among attested and unattested clusters.

In order to model the acceptability of the nonce words used in Scholes (1966), the feature-based model was trained and tested in two different ways. First, it was trained on the same corpus of onsets employed by Hayes & Wilson and tested on the experimental onset clusters.[10] As with all of the models compared by Hayes & Wilson, this model's predictions are based solely on the onset cluster, and ignore any differences in the remainder of the word. Second, in order to probe the possible effect of the varying rhymes in the Scholes (1966) items, the model was trained on the entire set of lemmas that occur in CELEX with non-zero frequency, calculating probabilities based on whole words. In this case, the model was tested on its predictions for the entire nonce word ([kɹʌn], [stɪn], [slɨk], [ʒpeɪl], etc.). In both cases, the model's predictions are based on type frequency of the sequences involved, while ignoring their token frequency. (Attempts to incorporate token frequency yielded somewhat worse results, in line with a similar finding by Hayes & Wilson; see also

[9] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[10] Scholes reports that four clusters were easily misperceived: [tl], [dl], [nl] and [vʒ]. These four items were excluded from all analyses. Careful inspection of the responses suggests that some other items may also have been sporadically misperceived; in particular, several clusters beginning with /f/ were accepted surprisingly often – e.g. [#fp] and [#fʃ] were accepted approximately as often as [#ml] and [#ʃn]. These responses diverge considerably from my own (adult native English) intuitions, and may indicate that the fricative was misperceived as /θ/ or even /s/ in some cases.

Bybee 1995, Pierrehumbert 2001, Albright 2002a, Albright & Hayes 2003, Buchwald 2007.) Finally, in order to make the results graphically comparable with Hayes & Wilson's predicted values, the predictions of the two models were then transformed to better approximate a linear fit. For the model trained on onsets only, the model's scores were transformed (following Hayes & Wilson) by a function of the form $y = x^{1/T}$, where the value of T was found by fitting (T = 2·01). For the model trained on whole words from CELEX, the optimal transformation turned out to be the log of the predicted probability.

The results, seen in Fig. 1, show that at least for purposes of modelling this data set, training on onsets alone is better at distinguishing among onset clusters – that is, it is advantageous to focus on the onset cluster and ignore the rest of the word. Because the acceptability scores have many ties, and because both the predictions and the ratings are non-normally distributed, Kendall's tau-c was calculated as a non-parametric measure of association (Kendall 1990). Consistent with the figures, the comparison shows that the onset-only model ($\tau_c = 0{\cdot}602$) captures subjects' preferences substantially better than the whole-word model ($\tau_c = 0{\cdot}417$).

One potential problem with the whole-word model is that it assigns equal weight to all of the substrings of the word: onset cluster, onset–nucleus transition and nucleus–coda transition. It has often been suggested that phonotactic constraints hold most strongly within the onset and within the rhyme, either as a universal principle or specifically for English (e.g. Selkirk 1982, Kessler & Treiman 1997). If this is correct, then the error of the whole-word model could be that it takes into account the onset–rhyme transition. In order to test this, whole-word scores for the test items were recalculated excluding the onset–rhyme transition. The result is an improvement over the model that includes every transition within the word ($\tau_c = 0{\cdot}508$), but still a worse match than the model that considers only the onset clusters.

There are several possible interpretations of this result. First, it might indicate, consistent with a key claim of Optimality Theory, that words are evaluated according to their worst/least probable constraint violation (in this case, the unattested onset cluster), with violations of lower-ranked constraints (here, the legal rhymes) contributing nothing to the evaluation. Although this is consistent with the current data, previous studies have shown that acceptability does sometimes depend on probabilities across the entire word (Coleman & Pierrehumbert 1997, Frisch *et al.* 2000 and §3.2 below). A second, more likely interpretation is that the importance of onset clusters in this particular data set is a by-product of the task, caused by the fact that the test items involved a wide variety of legal and illegal onset clusters, but only a small number of mostly common rhymes. This may have had the effect of making the onset clusters highly salient, and focusing subjects' attention on them (Sendlmeier 1987). Thus, we may conclude (in agreement with Hayes & Wilson 2008) that it is appropriate in this case to evaluate the statistical model based on its performance using a training corpus consisting solely of word onsets.

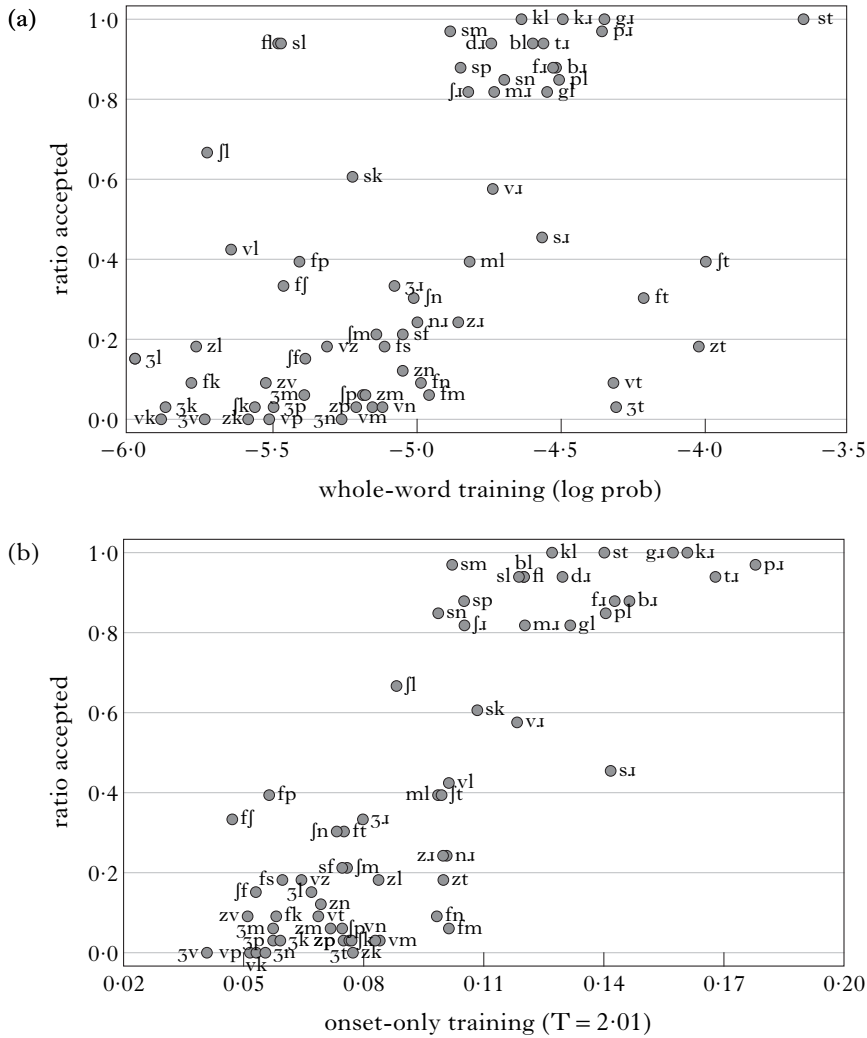


*Figure 1*
Model predictions for Scholes (1966) onset-cluster acceptability.
(a) Whole-word training; (b) onset-only training.

The fact that the model makes headway in predicting preferences among novel onset clusters over others bolsters the claim that feature-based generalisation is useful in modelling the acceptability of unattested sequences. We are now in a position to ask how this particular strategy of assessing similarity to existing clusters compares with Hayes & Wilson's constraint-based approach. Overall, the predictions of the Hayes &

| model | attested clusters | unattested clusters |
|---|---|---|
| feature-based bigrams | 0·291 | 0·355 |
| Hayes & Wilson (2008) | 0·166 | 0·528 |
| segment-based bigrams | 0·343 | undefined |
| Vitevitch & Luce (2004) | 0·343 | undefined |

*Table III*
Model comparison on attested and unattested clusters (Kendall's $\tau_c$).

Wilson model[11] show an impressively strong correlation with Scholes' observed acceptance rates: Kendall's $\tau_c = 0{\cdot}702$.[12] This success is based on two features of the model's predictions. First, like humans, the model prefers nearly all attested clusters over unattested clusters. In addition, among unattested clusters, the model often prefers English-like clusters such as [#ʃl] or [#vɹ] over completely un-English clusters such [#fm] or [#vk]. Crucially, however, the Hayes & Wilson model predicts almost no differences among attested clusters such as [#pɹ] and [#sk], even though these items differ considerably in their judged acceptability.

As can be seen in the upper portion of Fig. 1b, the feature-based biphone model predicts substantial differences among attested clusters – sometimes correctly, and sometimes not. This qualitative difference between the two models can be seen in Table III, which compares performance on attested and unattested clusters separately.[13] We see that although the feature-based biphone model does not achieve quite as high a $\tau_c$ value as the best models for either type of cluster, it displays consistent and moderately strong performance across both types of test items. By contrast, the Hayes & Wilson model does substantially better for unattested clusters and worse for attested clusters; in fact, the score of 0·166 for attested clusters is based entirely on the fact that the model correctly assigns a lower score to [#ʃɹ] than to most other clusters. Conversely, segment-based bigram models such as a simple biphone model or the positional bigram model of Vitevitch & Luce (2004) do quite well in modelling differences among attested clusters, but are fundamentally unable to distinguish among unattested clusters. As discussed above, the

[11] Downloaded (August 2008) from http://www.linguistics.ucla.edu/people/hayes/phonotactics/index.htm.

[12] Hayes & Wilson employ Spearman's $\rho$ rather than Kendall's $\tau$, but the comparison yields completely equivalent results: their model achieves $\rho(60) = 0{\cdot}889$, while the current model achieves $\rho(60) = 0{\cdot}793$ when trained on onsets only (adapted CMU), $\rho(60) = 0{\cdot}551$ when trained on whole words (CELEX lemmas) and $\rho(60) = 0{\cdot}684$ when onset–nucleus transitions are ignored (CELEX lemmas). We do not consider here Pearson correlation ($r$) values, since they are likely to be inflated due to the fact that all models under consideration produce many tied predictions at or near the endpoints of the scale.

[13] The cluster [#sf] is excluded from this analysis, since it had been removed from the onset-training corpus. It was thus unattested for the models, but potentially attested for Scholes' subjects (*sphere*, *sphynx*, etc.).

feature-based bigram model tends to approximate segmental bigram models for attested clusters, while using features to estimate the likelihood of unattested clusters. The results do not appear to be quite as good as either mode of generalisation alone. However, the precise numerical magnitude of the difference between the models should be interpreted with caution. In order to see why, it is useful to consider separately the performance of the feature-based model in comparison with the Hayes & Wilson model for unattested clusters, and in comparison with a segmental bigram model for attested clusters.

As a point of comparison for performance on unattested clusters, the difference between the feature-based bigram model ($\tau_c = 0{\cdot}355$) and Hayes & Wilson's (2008) constraint-based model ($\tau_c = 0{\cdot}528$) is shown visually in Fig. 2. We see by visual inspection that the Hayes & Wilson model does achieve a better linear fit, but as with the segmental biphone model, this is accomplished in part by making fewer distinctions. In order to confirm that the difference between $\tau_c$ of 0·355 and 0·528 is not a qualitative difference in the way that the difference between 0·166 and 0·528 is, separate logistic regressions were carried out, using the feature-based bigram model and the Hayes & Wilson constraint model to predict the proportion of 'yes' responses (out of 33) for attested and unattested clusters. The results show that the feature-based bigram model is a highly significant predictor of acceptability for both attested ($p < 0{\cdot}001$) and unattested ($p < 0{\cdot}0001$) clusters, while the Hayes & Wilson model is a significant predictor only of unattested ($p < 0{\cdot}0001$) but not of attested ($p = 0{\cdot}08$) clusters. We may therefore conclude that although the feature-based model is far from perfect in its predictions, it provides a reasonable first pass estimate of the acceptability of novel clusters.

Turning to performance on attested clusters, Table III shows that here, too, the feature-based model does not achieve quite as high a degree of success as the best model, which in this case is the segment-based biphone model. It was suggested above that the feature-based biphone model performs as well as it does for attested sequences because the instantiation costs favour selecting the most specific possible natural classes (i.e. individual segments), which has the effect of approximating a segment-based biphone model. The comparison in Table III shows that the models' predictions are not identical, however, and, at least as measured by Kendall's $\tau_c$, the segment-based model still does slightly better. It is tempting to infer on the basis of this comparison that attested sequences are actually evaluated in terms of segmental probability, and that feature-based generalisation is employed only in cases of unattested sequences. This conclusion is premature, however, since it appears that for these items, the predictions of the feature-based model are actually approximately equivalent to or even better than those of the segmental model. In order to compare the segmental and feature-based models more directly, a logistic regression was carried out in which both models were used to jointly predict the proportion of 'yes' responses (out of 33) for attested onset clusters. The results show that either model alone is a significant

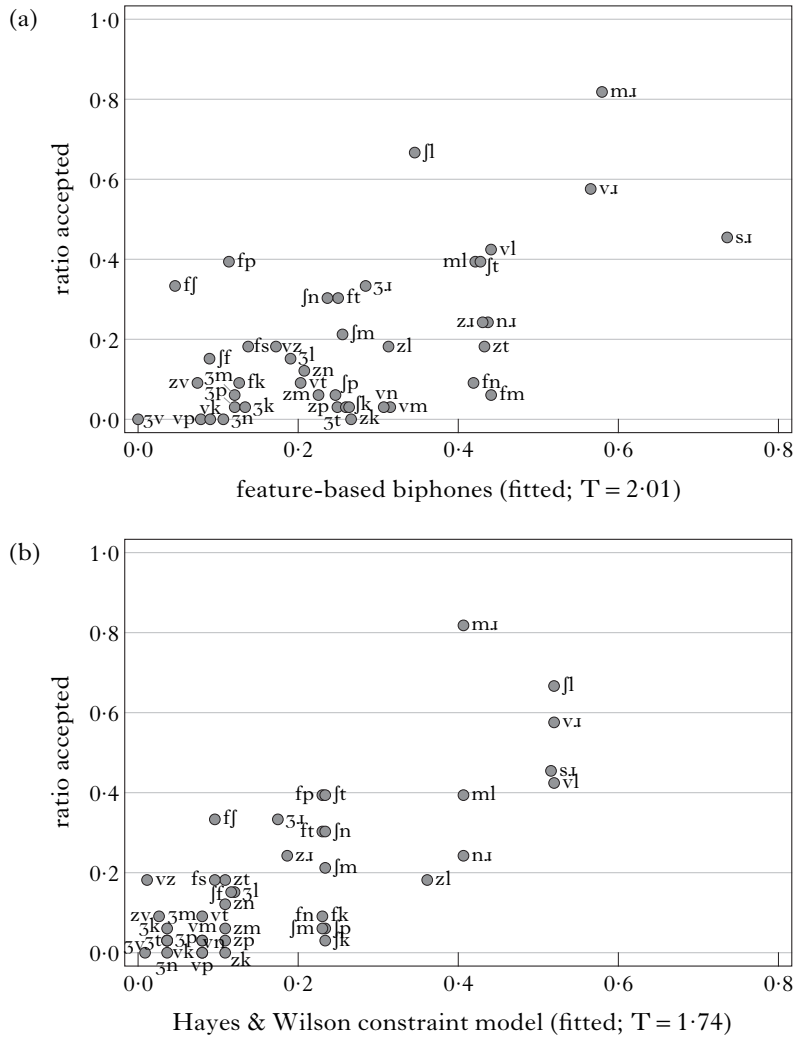


*Figure 2*

Generalisation to unattested clusters. (a) Feature-based biphones; (b) Hayes & Wilson constraint model.

predictor of acceptance rates (segments: $p < 0{\cdot}05$; features: $p < 0{\cdot}001$). However, when the two models are considered together, only the feature-based model is indispensable: the feature-based model is still added significantly ($p < 0{\cdot}05$) even after the segmental model is taken into account, but the converse is not true. By this criterion, the feature-based biphone model is not merely a poor approximation of segmental bigrams, but is actually a better model of how speakers evaluate the acceptability of not just unattested, but also attested combinations of segments. We must

interpret this result with caution, since the Kendall's $\tau_c$ values in Table III suggest a slight advantage for the segmental biphone model instead. In §3.2, we will see evidence that neither the segment-based nor the feature-based model can be fully reduced to the other; this point will be discussed further in §4.1.

The upshot of these comparisons is that the feature-based biphone model is not merely a second-best compromise, but actually appears to be roughly comparable to (and in some respects even better than) the models that achieve the highest Kendall's $\tau_c$ for attested and unattested clusters respectively. This supports the conclusion that a model employing features and natural classes provides a reasonable approximation of how subjects generalise to novel items involving both attested and unattested sequences, with a comparable balance of performance on both. This can be contrasted with the segmental biphone model, which fails to distinguish among unattested clusters, and the Hayes & Wilson model, which does not adequately distinguish among attested clusters. It should be emphasised that the lack of differentiation among attested clusters is not an intrinsic feature of the Hayes & Wilson (2008) model, but is a consequence of the fact that the model is biased towards generality and against overfitting by two mechanisms: a greedy constraint-discovery procedure that seeks to discover the most general constraints first, and a bias not to use extra constraints (i.e. to keep all weights at zero). It is left as a matter for future research to determine whether a less biased weighting objective or a different strategy for constraint induction would allow the constraint-based model to better capture differences among attested sequences by using constraints that are featurally more specific, while retaining strong performance on unattested clusters.

### 3.2 Monosyllabic nonce words

In the preceding section, the model's ability to capture gradient preferences was tested in a controlled comparison among items that differ in (primarily) one respect: the distance between their onset clusters and the set of attested clusters. A different use of phonotactic models in the literature has been to model and control for the overall acceptability of diverse words based on the frequency of their subparts (Vitevitch *et al.* 1996, Coleman & Pierrehumbert 1997, Frisch *et al.* 2000, Bailey & Hahn 2001, Shademan 2007). In this section, we test the model on acceptability ratings of a more diverse set of monosyllabic nonce words. Unlike the Scholes (1966) experiment, in which it is plausible to assume that subjects were paying attention primarily to the varying onsets, this task provided no such cue that would have encouraged subjects to focus on a particular aspect of the word.

The data set consists of acceptability ratings for a set of 88 non-words collected by Albright & Hayes (2003) as norming data for a past tense study.[14] The set of test items were chosen to cover a range of phonotactic

[14] Available (March 2009) for download from http://www.mit.edu/~albright.

plausibility, ranging from frequent onsets and rhymes (e.g. *drit* [dɹɪt], *stire* [staɪɹ], *pank* [pæŋk], *fleep* [fliːp], *blafe* [bleɪf]) to rare or phonotactically marginal sequences (e.g. [pwʌdz] (*[pw]), [θɹɔɪks] (*[ɔɪk]), [ʃwuːʒ] (*[ʃw]), [skɪk], [snʌm] (*$sC_1VC_1$, *sNVN)). A minority of the nonce words contained completely unattested rhymes (e.g. *smairg* [smɛɹg], *smeelth* [smiːlθ]).[15] The nonce words were presented auditorily in random order, as verbs in a carrier sentence ('*Blafe*. I like to *blafe*.'). Participants repeated each 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 the analysis. One very un-English item (*bzarshk* [bzaɹʃk]) was used during training as an example of a word that most English speakers felt could not be a possible word; this was contrasted with [kɪp], which was offered as a word that speakers would generally find very acceptable.[16] Nineteen native speakers of American English were paid to take part in the ratings task, which lasted approximately ten minutes.

A preliminary issue that should be addressed is to what extent ratings of diverse nonce words can be meaningfully compared with one another. I argued above that the comparisons in Scholes (1966) may have been performed by focusing on the relative acceptability of just the word-initial clusters: [ʃtɪn] ≻ [vtɪn]. The ratings in this task, on the other hand, required subjects to distinguish among items that differ in several respects at once – e.g. *trisk* [tɹɪsk] *vs*. *nace* [neɪs] *vs*. *bredge* [bɹɛʤ]. Previous studies have often assumed that responses along a single scale straightforwardly reflect differences along an absolute cline of acceptability (Greenberg & Jenkins 1964, Ohala & Ohala 1986, Coleman & Pierrehumbert 1997, Frisch *et al*. 2000, Bailey & Hahn 2001). However, as Bailey & Hahn (2001) note, data concerning relative preferences among diverse words can be quite noisy and detailed differences may be difficult to interpret. For example, the results of the current study show a preference in mean ratings for *spack* [spæk] (5·2) over *teep* [tiːp] (4·6) over *nold* [noʊld] (4·0). However, unlike more clear-cut comparisons such as *teep* ≻ *thweeks* (2·5), these differences are not so readily confirmed by native speakers through informal introspection ('would *teep* be more possible than *nold* as a word of English?'). Thus it is reasonable to question whether the numerical ratings truly reflect subtle preferences such as *teep* ≻ *nold*, or whether

[15] Some of these items may be familiar to readers as 'distant pseudo-regular' verbs from Prasada & Pinker's (1993) study of English past tenses. The nonce items constructed by Albright & Hayes also included three items that are real words, but not verbs: [ʃiː] *she*, [fɹoʊ] *fro* and [ɹaɪf] *rife*. These items were excluded from the current analyses.

[16] The word *kip* is in fact a lexical item in some dialects or professions (and is a unit of currency of Laos), but it was not generally familiar to the subjects of the study, who were mostly speakers of California English. The *Oxford English Dictionary* lists no clear uses of *kip* as a verb since the fifteenth century.

speakers actually make much coarser distinctions, with small differences largely reflecting experimental noise.

There are two reasons to believe that the differences in ratings of entire words are nonetheless reliable, and may be assigned a meaningful interpretation. First, it appears that in spite of the lack of clear intuitive differences between many of the items, participants in this experiment show a relatively high degree of between-subject agreement, and the differences in mean ratings are replicable. For example, the correlation between participants 1–10 *vs.* 11–19 for the 88 test items was $r(86) = 0{\cdot}847$, while the $\omega^2$ value for all 19 subjects was 0·45, indicating substantial word-by-word differences relative to noise.[17] These results suggest that, at least for the current set of items and experimental set-up, relative preferences among diverse nonce words are robust.

The second and more important source of support for a unified scale of acceptability comes from the fact that the observed differences are not only replicable, but also largely interpretable in terms of phonotactic probability. As a preliminary check of the relation between sequence frequency and acceptability, a simple segment-based biphone model was trained on the set of lemmas that occur in CELEX with frequency greater than zero. The model was used to derive predicted values for Albright & Hayes' (2003) nonce items, calculated as the product of the individual bigram transitional probabilities (including word boundaries). Figure 3a shows participants' mean ratings plotted as a function of their bigram probability. As can be seen, the model does a reasonably good job in predicting speakers' relative preferences (Kendall's $\tau = 0{\cdot}474$, $p < 0{\cdot}0001$), including preferences such as *spack* (log prob = −12·70) ≻ *teep* (−13·32) ≻ *nold* (−14·68). This result is not especially surprising, since it echoes findings such as those of Frisch *et al.* (2000) that ratings of entire words may closely track the combined probabilities of their subparts. However, it is nonetheless useful, since it provides reason to believe that the ratings in this task are in fact determined to a large extent by phonotactic probability, as opposed to factors such as the inferred meaning of the word or the meaning of phonologically similar words (Ramscar & Yarlett 2003). In addition, it shows that at least for this data, the acceptability of a nonce word is appropriately modelled as the joint probability of its subparts.[18]

Against this baseline, it is now possible to consider the evidence for a model that employs representations with more linguistic structure. Whereas Frisch *et al.* (2000) find that for words of varying lengths, it is advantageous to consider the probability of syllabic constituents larger

[17] As a replication, ratings for 77 of these items were also collected from 20 additional native English-speaking subjects as part of a follow-up study, in which nonce items were presented as both nouns and verbs (counterbalanced across subjects). The correlation between the mean ratings from this second group and the 19 subjects in the original experiment is comparable: $r(75) = 0{\cdot}853$.

[18] By contrast, the phonotactic probability calculator of Vitevitch & Luce (2004), which relies on averaged positional biphone probabilities across the word, does not do nearly as well as the simple biphone model in capturing preferences among these nonce items (Kendall's $\tau = 0{\cdot}177$, $p = 0{\cdot}016$).

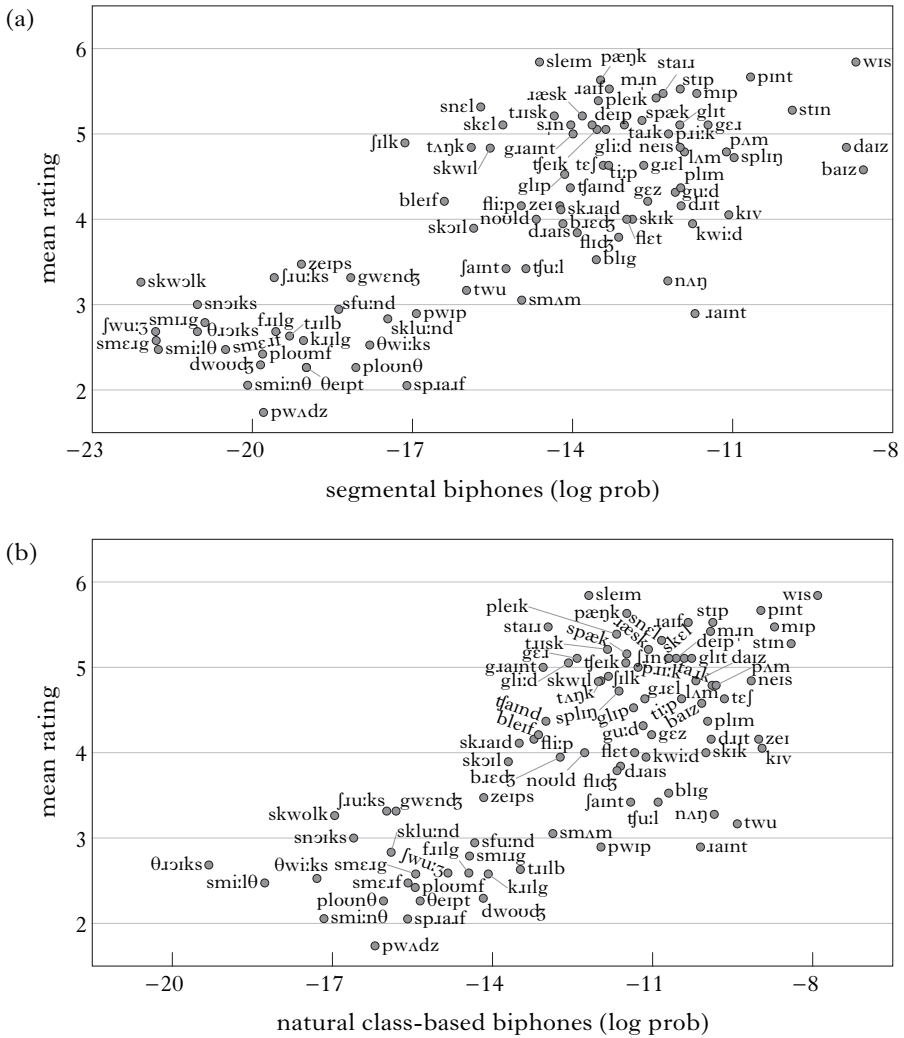


*Figure 3*

Biphone models of whole-word ratings.
(a) Segmental biphones; (b) natural class-based biphones.

than the segment (onsets, rhymes), in the present case, many test items have onsets or rhymes that are rare ([#sf, #skl, uːʒ#]) or unattested ([#ʃw, ɪlb#, ɔɪks#]), and are thus not differentiated by frequency of the constituent taken as a whole. For such comparisons, we would expect the model to benefit from an ability to evaluate rare or unattested sequences by considering their similarity to other attested combinations – precisely as the natural class-based model is designed to do. In order to investigate the possibility of modelling nonce word acceptability in terms of

sequences of natural classes, the model in §2 was likewise trained on the set of CELEX lemmas with frequency greater than zero and used to derive predictions for the set of 88 nonce words. The results, in Fig. 3b, show that the class-based model also does a reasonable job in predicting speakers' preferences (Kendall's $\tau = 0{\cdot}453$, $p < 0{\cdot}0001$). As discussed above, part of the reason for this similar success is due to the fact that the natural-class based model closely mirrors segmental *n*-gram probability for attested combinations of segments. However, comparison of Figs 3a and b reveal that the models are not completely isomorphic; there are some cases in which natural class-based reasoning is able to find support that goes beyond the literal biphones (e.g. *shilk* [ʃɪlk]), and other cases in which class-based parsing puts items at a relative disadvantage (e.g. *dize* [daɪz], *bize* [baɪz]). The question, then, is whether natural class-based reasoning constitutes a significant improvement over purely segmental counts for modelling items composed of attested biphones.

Unfortunately, by several simple criteria, the answer is 'no'. Comparing the Kendall's $\tau$ values reported above, for example, we see slightly better performance for segmental biphones than for natural class-based biphones (4·74 *vs.* 4·53); Spearman correlations show a similar result ($\rho(86) = 0{\cdot}673$ *vs.* 0·636). In order to assess whether this difference is significant, individual participant ratings were first transformed with a standard arcsine transformation, employing a small correction for ratings at the extreme endpoints of the scale (see Sheskin 2004: 408). The transformed subject ratings were then fit with linear mixed-effect models, using the relativised maximum likelihood technique provided by the lme4 package (Bates *et al.* 2008) of R (R Development Core Team 2008). The linear models treated segmental and natural class biphone probability as fixed effects, and subject identity (and any possible interaction with subject) as a random effect. A likelihood ratio test was used to compare the predictions of models using segmental *vs.* natural class-based biphones. The results show that segmental biphones actually yield significantly better predictions ($p < 0{\cdot}0001$), in accordance with the informal comparison of $\tau$ values above.[19] Thus it does not appear that allowing the model to refer to natural classes is a straightforward improvement on the segmental model, though the precise advantage for one model over the other may depend on the ratio of common *vs.* rare/unattested sequences in the set of test items.

Even if the feature-based model cannot subsume the segmental model, we may ask whether it nonetheless makes an independent contribution. In order to test this, two nested linear mixed effect models were fit: one using segmental biphones to predict subject ratings, and one using both

[19] Pinheiro & Bates (2000) point out that this strategy for comparing fixed effects may be overly liberal, though such problems are unlikely in this case because there are many observations and few degrees of freedom. Even if the comparison turns out to be non-significant, however, we would still find no support for the idea that the natural class-based model outperforms simple segmental biphones.

segmental and natural class-based biphones. (In both cases, subject identity was included as a random factor.) Here, the likelihood ratio test compares the value of adding natural class-based biphones to a model that already includes segmental biphones. The result reveals a significant positive contribution ($\chi^2 = 12{\cdot}443$, $p = 0{\cdot}014$). Although the degree of improvement is small, it has been argued that such comparisons of nested models tend to be overly conservative (i.e. p may actually be lower, or more strongly significant; Pinheiro & Bates 2000: 87). We therefore find support for the idea that natural class-based reasoning may be helpful not just for unattested sequences, but also in modelling generalisation to attested combinations of segments.

A question that arises is whether speakers truly generalise based on natural classes, or whether they might generalise based on some other mechanism such as segmental similarity. Concretely, suppose that speakers find sequences like [#sɹ] or [lg#] moderately acceptable not because they embody attested combinations of natural classes ([+strident] [+rhotic] and [+lateral][−sonorant, −continuant, +voice] respectively), but rather because they are perceptually similar to specific members of those classes ([#ʃɹ, ld#]). Hayes & Wilson (2008) test this question using Bailey & Hahn's (2001) Generalised Neighbourhood Model (GNM). In this model, nonce item are aligned with known words according to the phonetic similarity of the segments in question, so that nonce clusters find greater support from clusters involving similar sounds. Hayes & Wilson find that for the task of modelling novel onset clusters, similarity to existing clusters is not as predictive as generalisation by natural classes. It is reasonable to ask whether the same holds for rare but attested sequences – e.g. does *shilk* [ʃɪlk] find support from its similarity to *silk*, *silt*, *shalt* and so on? In order to test this, a version of the GNM was implemented, using segmental similarity values generously provided by Todd Bailey. The model was trained on the entire set of CELEX word forms,[20] and used to derive predictions for the 88 test items. As before, nested linear mixed-effect models were fit with subjects as a random effect and the predictions of the GNM and segmental biphones as fixed effects, with and without the predictions of natural class-based biphones as an additional fixed effect. The goal of this comparison is to determine whether there is a significant benefit to incorporating the predictions of the natural class model. A likelihood ratio test reveals that the natural class model still provides a unique independent contribution – in fact, one that is now highly significant ($\chi^2 = 27{\cdot}27$, $p < 0{\cdot}0001$). We therefore conclude that the effect of generalisation by natural classes is an independent effect

[20] Comparison of models based on lemmas alone or on entire word forms revealed that the latter performed slightly better for this set of test items. The set of CELEX word forms was processed to ignore entries for homophonous or duplicate word forms. The Generalised Neighbourhood Model, like its progenitor the Generalised Context Model (Nosofsky 1986), also has a number of free parameters that must be set by fitting. In the current case these were: insertion/deletion = 0·7, D = 5·75, with no influence of token frequency (A = 0, B = 0, C = 1).

which cannot be attributed to an independent analogical effect based on similarity to known items.

These results can be contrasted with those of §3.1, in which the predictions of the feature-based biphone model were seen to overlap with and perhaps even subsume the segment-based model. The task of modelling entire words provides greater opportunity for the predictions of the two models to diverge for two reasons: first, it involves much larger and phonologically more diverse set of biphones, and second, the set of test items in question includes sequences with a wider range of frequencies. Under these conditions, we find that the two models do in fact make distinct and complementary predictions. Given this result, we may return to the hypothesis put forward in the previous section, that natural class-based generalisation is employed primarily in case of unattested or rare phoneme combinations. In order to test this, the (subject-specific) error of the segment-based and natural class-based biphone models was calculated (= the fitted residuals), and the average error of each model for each verb was determined. If it is correct that speakers use natural classes only when the segment combination is rare or unattested, then we would expect the natural class-based model to be more accurate than the biphone model (i.e. have smaller error) for items with rare combinations such as [smiːlθ] or [snɔɪks], while the segment-based model would be more accurate for items with frequently attested combinations such as [stɪn] or [tiːp]. Contrary to this prediction, it turns out that the error of both models is distributed fairly evenly across the range of items. This suggests that segment-based and natural-class based generalisation are not employed in complementary fashion for different types of words.

In summary, the results of this section support the claim that speakers generalise using phonological features when assessing the acceptability of nonce words, even in cases where the sequences involved are attested and frequent. This result is complicated, however, by the fact that we find evidence for distinct effects of segment-based and natural-class based biphone scores. In §4.1, we consider the question of how these separate effects may arise.

## 4 Discussion

The preceding sections provide two types of evidence in support of the model presented in §2, which generalises to nonce items by analysing them in terms of phonological structure (features). When we consider performance on the Scholes (1966) onset cluster data, we see that the model provides a unified account of generalisation to both attested and unattested clusters, combining many of the advantages of a simple segmental biphone model for attested sequences and the Hayes & Wilson (2008) constraint-induction model for unattested sequences. When we consider entire nonce monosyllables such as those tested by Albright & Hayes (2003), we find that here, too, the model makes substantial headway in predicting acceptability ratings (Fig. 3b), and provides a significant

unique contribution even when simpler segment-based biphones and neighbourhood effects are taken into account. At the same time, it appears that the model's predictions cannot wholly supplant those of the simpler models. Thus, we are compelled to consider where such a model may fit in the broader phonological system. We take on this question in §4.1, considering what aspect of phonological performance or competence this model may reflect, and how it relates to grammar. §4.2 compares the current results to those of a similar study by Bailey & Hahn (2001), highlighting some discrepancies and suggesting some reasons for the differences. Finally, §4.3 briefly highlights one respect in which the current results converge with several other studies of gradient acceptability: the importance of type frequency.

## 4.1 Natural class effects and phonological grammar

In §3.1, it was argued that an advantage of the way in which the proposed model evaluates natural classes is that it tends to characterise frequently attested sequences as specifically as possible – i.e. as biphones of individual segments. The fact that the two models overlap so closely for attested sequences might lead us to hypothesise that the feature-based model is simply a more general version of the simpler segment-based model, equipped to distinguish among unattested as well as attested sequences. The results of §3.2 reveal that this is not the case, however; in fact, the two models appear to make distinct predictions, both of which play a significant role in capturing acceptability ratings.

One possible interpretation that cannot be ruled out is that neither model is correct, and that we should seek a unified model that incorporates aspects of both. For example, if the feature-based model were modified to change how it assesses instantiation costs, it might mirror even more closely the segmental model in cases where this is helpful, while retaining an ability to use feature-based generalisation in cases where humans do. The residuals analysis at the end of §3.2 suggests that an easy solution along these lines is unlikely, since the unique positive contributions of the feature-based model are not confined to rare or unattested sequences. Instead, it appears that the two models simply represent different ways of evaluating the likelihood of sequences, both of which are reflected in human acceptability ratings.

This conclusion may seem undesirable, since it requires two overlapping models to model a single type of response. However, as discussed in the introduction, phonotactic effects are often thought to originate via two distinct mechanisms: an initial stage of processing that evaluates phoneme combinations for purposes of morpheme segmentation and lexical access, and a phonological grammar that can be invoked to evaluate the well-formedness of words. Clearly, the most economical model of phonological knowledge would be one in which both tasks relied on a single set of grammatical constraints. However, it is entirely possible (and indeed, it is often assumed; e.g. Auer & Luce 2005) that the early stages of processing

involve a coarser evaluation in which the details of competing phonological parses play less of a role. If we accept this possibility, then the current results can be interpreted straightforwardly as the result of two stages of evaluation: a preliminary initial assessment of the probability of segment combinations, and a subsequent grammatical evaluation of the likelihood of featural co-occurrences.

If this interpretation is correct, then it implies that gradient distinctions among attested sequences are not solely a by-product of the early stages of processing, but rather arise though a probabilistic grammatical evaluation in terms of competing phonological structures. This is parallel to claims by Coleman & Pierrehumbert (1997), Frisch *et al.* (2000) and Hay *et al.* (2004), who argue that the acceptability of word-medial sequences is best modelled in terms of competing parses consistent with a grammar of possible syllable structures. The current result strengthens the conclusion that a distinct gradient grammatical evaluation is necessary, since we are able to identify separate and unique contributions of both the simpler evaluation (involving a rougher parse into segments) and the detailed and costlier parse into most likely natural class combinations.

### 4.2 Lexical *vs.* phonotactic knowledge

In addition to the interplay of segment-based and feature-based probabilities, the results of §3.2 also show an interaction with another important factor influencing acceptability intuitions: neighbourhood effects based on similarity to existing lexical items. The simultaneous existence of phonotactic and lexical influences is by no means a new finding; for example, in a detailed comparison of the role of different phonotactic and lexical predictors on nonce word ratings, Bailey & Hahn (2001) find that both make significant independent contributions. There appears to be an important quantitative difference between Bailey & Hahn's results and those of §3.2, however: whereas Bailey & Hahn find a relatively strong effect of neighbourhood density (as measured by their Generalised Neighbourhood Model) and only a modest effect of phonotactic probability (as measured by segment-based biphone and triphone models), the current results show the opposite asymmetry. This can be seen in several different ways. First, when we examine the linear mixed effect models from §3.2 which included the GNM along with the segmental and feature-based biphone models, we find that the GNM does not make a significant unique contribution ($t = 1{\cdot}093$, $p = 0{\cdot}275$).[21] Furthermore, when we compare nested models that include both biphone models and differ only in the inclusion of the GNM, a likelihood ratio test reveals that adding the GNM does not yield a significantly more accurate model ($\chi^2 = 3{\cdot}89$, $p = 0{\cdot}569$). Thus, in contrast to Bailey & Hahn's results, we find that the Albright &

[21] As discussed by Baayen (2008: 269–270), the assessment of significance values for such mixed-effect models is controversial. Since the current model is based on many observations, we follow Baayen's suggestion that the number of observations minus the number of fixed effects is an appropriate approximation.

Hayes (2003) ratings appear to be based primarily on phonotactic probability.

Why would different experiments elicit larger or smaller lexical effects? Shademan (2006) argues that this may be a task-dependent difference, caused by the fact that Bailey & Hahn included real word fillers among their test items. Shademan hypothesises 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 neighbours can indeed be modulated by whether or not real word items are included in the task. If this is right, then the Albright & Hayes experiment (which did not include real word fillers) would favour evaluation with the phonotactic grammar.[22]

Another factor that may contribute to the discrepancy involves a difference in how the nonce words were constructed. Bailey & Hahn (2001) tested items composed primarily of attested sequences that are grammatical according to standard analyses of English phonotactics (e.g. Clements & Keyser 1983). The Albright & Hayes nonce words, on the other hand, contained a larger number of rare or marginal sequences. Since the GMN evaluates words based on their overall similarity to existing words, it is possible for items with marginal sequences to nonetheless gain strong support from existing words that have significant overlap elsewhere in the word. To take an extreme example, hypothetical [bzɛkfəst] may receive strong support based on its similarity to [bɹɛkfəst] *breakfast*, without any particular penalty for the illegal [#bz] cluster. More generally, words with rare or illegal sequences may be quite similar to existing words in other respects, leading the GNM to predict that they will have high scores.

A 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' mean ratings, and calculating the residuals. The non-words with the greatest positive residuals, or those which the GNM most seriously overestimated, are shown in (7).

(7) *Non-words for which the GNM most seriously overestimated goodness*

| | | | | | |
|---|---|---|---|---|---|
| throiks | [θɹɔɪks] | smeelth | [smiːlθ] | dize | [daɪz] |
| shwouge | [ʃwuːʒ] | smairf | [smɛɹf] | thaped | [θeɪpt] |
| rin't | [ɹaɪnt] | thweeks | [θwiːks] | smeenth | [smiːnθ] |
| frilg | [fɹɪlg] | ploamph | [ploʊmf] | sprarf | [spɹaɹf] |
| krilg | [kɹɪlg] | dwoge | [dwoʊʤ] | bize | [baɪz] |
| smairg | [smɛɹg] | ploanth | [ploʊnθ] | pwuds | [pwʌdz] |
| trilb | [tɹɪlb] | | | | |

[22] As observed in notes 15 and 16, the Albright & Hayes items did include three items that exist as words, but not as verbs. 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.

The majority of these (84%) items contain some sort of phonotactic violation. Eight of them contain a sequence that would be fairly uncontroversially ruled out by standard analyses of English phonotactics, while eight of them contain rhymes that are of dubious legality.[23] The GNM has no mechanism for attending to or encoding such violations. Moreover, since neither Bailey & Hahn nor Albright & Hayes included significant numbers of words with illegal sequences, data from these studies allow us to probe only a tiny piece of the overall picture of how well the models account for gradient acceptability. It seems likely that if more illegal sequences had been included, embodying a broader range of more serious phonotactic violations, phonotactic evaluation would be an even stronger determinant of acceptability ratings.

It is also instructive to examine the non-words which the GNM most seriously underestimated, listed in (8). Unlike the items in (7), none of these items in (8) contains illegal or unattested sequences. This asymmetry is significant, since it confirms that the model is systematically overestimating the acceptability of words with phonotactic violations. Instead, the items in (8) happen to be somewhat isolated in the lexicon. Just as subjects in the Albright & Hayes (2003) study were evidently able to focus on phonotactic violations in spite of similarity to existing words, it appears that they were also not overly bothered by a lack of many similar existing words in the case of legal sequences.

(8) *Non-words for which the GNM most seriously underestimated goodness*

| | | | | | |
|---|---|---|---|---|---|
| slame | [sleɪm] | plake | [pleɪk] | shilk | [ʃɪlk] |
| stire | [staɪɹ] | mip | [mɪp] | squill | [skwɪl] |
| pank | [pæŋk] | wiss | [wɪs] | gare | [gɛɹ] |
| snell | [snɛl] | grin't | [gɹaɪnt] | preek | [pɹiːk] |
| rask | [ɹæsk] | skell | [skɛl] | glit | [glɪt] |
| trisk | [tɹɪsk] | spack | [spæk] | murn | [mɹ̩n] |
| stip | [stɪp] | stin | [stɪn] | | |

Crucially, a parallel error analysis of the feature-based model reveals no such systematic difference between words that the model most severely overestimates *vs.* underestimates.

The upshot of this analysis is that phonotactic effects in ratings of nonce words are not always small and highly confounded with neighbourhood effects, as they are in Bailey & Hahn's (2001) study. This result is partly of practical importance, since it confirms that such tasks may indeed be an informative source of data concerning gradient phonotactic grammar. It is also of theoretical importance, since it constitutes additional evidence in

[23] The items with clearly illegal sequences include *pwuds* (labial dissimilation; Hammond 1999: 56), *shwouge* (*[ʃC]; Hammond 1999: 101), *thweeks*, *throiks*, *thaped*, *ploamph*, *smairg* (*tense vowel + non-coronal cluster; Selkirk 1982) and *sprarf* (*[CɹVɹ]; see Davis 1984 for discussion of the related *rVrC constraint). Dubious items include *frilg*, *krilg*, *trilb* ([lb#] very rare, [lg#] unattested), *ploanth*, *smeenth*, *smeelth* (tense vowel + [Cθ] unattested) and *dwoge* (rhyme [oʊʤ] unattested).

favour of a probabilistic phonotactic grammar distinct from the lexicon: the nonce words analysed in §3.2 should be mostly grammatical under a categorical account of English phonology, yet the best account of gradient differences among them is stated not in terms of neighbourhood support, but rather in terms of segment or feature combinations.

### 4.3 Token frequency

There is one respect in which the current results confirm a number of previous claims in the literature that pattern strength is determined by type, not token frequency (Bybee 1995, Albright 2002b, Albright & Hayes 2003, Hay *et al*. 2004, Buchwald 2007). None of the models reported here derived any advantage from weighting the contribution of individual data according to their token frequency. This is seen in several ways: (i) the Generalised Neighbourhood Model was found to perform best when the frequency term was set to zero; (ii) the natural class-based model did best when instantiation costs were calculated without reference to the relative frequency of different segments; and (iii) the Vitevitch & Luce model, which intrinsically takes token frequency into account, did not come out ahead by virtue of having this ability. This finding contrasts with that of Bailey & Hahn (2001), who found a small (≈1%) but significant contribution for frequency weighting. In fact, for most of the simulations reported here, taking token frequency into account makes very little difference in the predictions of the models, since most words in the lexicon are very low frequency and a boost for high token frequency words gives more influence to just a small set of items. When token frequency does make a difference, though, it tends to be a deleterious one. This fact may potentially support the idea that the acceptability ratings in the current studies reflect gradient grammatical knowledge and is not a by-product of online processing or lexical access, in which token frequency has a strong effect. At present this result is merely suggestive, however, since not all online tasks show effects of token frequency, while grammar, too, has sometimes been argued to be sensitive to token frequency. Furthermore, neither study included items designed to test specifically for the contribution of token-frequency effects, so the conclusions here must be based on tentative post hoc comparisons.

## 5 Conclusion

In this paper, I have presented and tested a model that employs phonological features and natural classes to predict phonotactic well-formedness. When the predictions of the model are compared against experimentally obtained acceptability ratings of nonce words, it emerges that the model is a strong predictor of acceptability of both attested and unattested sequences. Furthermore, it appears that the predictions of the model are distinct from those of segment-based biphone models, and

provide a unique independent contribution. The fact that we observe separate effects of segment-based and feature-based biphone probability may indicate two levels of evaluation, which could correspond to a shallow assessment of segment co-occurrence during the initial stages of perception, and a subsequent parsing into optimal combinations of phonological features as part of grammatical evaluation. Teasing apart these effects and testing the use and limits of feature-based generalisation in predicting well-formedness of a broader range of unattested sequences are left for future experimental and modelling work.

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

| *word* | *rating* | *word* | *rating* | *word* | *rating* | *word* | *rating* |
|---|---|---|---|---|---|---|---|
| wɪs | 5·84 | gliːd | 5·05 | dɹɪt | 4·16 | snɔɪks | 3·00 |
| sleɪm | 5·84 | pɹiːk | 5·00 | fliːp | 4·16 | sfuːnd | 2·94 |
| pɪnt | 5·67 | gɹaɪnt | 5·00 | skɹaɪd | 4·11 | pwɪp | 2·89 |
| pæŋk | 5·63 | ʃɪlk | 4·89 | kɪv | 4·05 | ɹaɪnt | 2·89 |
| ɹaɪf | 5·53 | daɪz | 4·84 | skɪk | 4·00 | skluːnd | 2·83 |
| stɪp | 5·53 | tʌŋk | 4·84 | flɛt | 4·00 | smiːɹg | 2·79 |
| mɪp | 5·47 | neɪs | 4·84 | noʊld | 4·00 | θɹɔɪks | 2·68 |
| staɪɹ | 5·47 | skwɪl | 4·83 | bɹɛdʒ | 3·95 | fɹɪlg | 2·68 |
| mɻn | 5·42 | lʌm | 4·79 | kwiːd | 3·95 | ʃwuːʒ | 2·68 |
| pleɪk | 5·39 | pʌm | 4·79 | skɔɪl | 3·89 | tɹɪlb | 2·63 |
| snɛl | 5·32 | splɪŋ | 4·72 | dɹaɪs | 3·84 | smɛɹg | 2·58 |
| stɪn | 5·28 | gɹɛl | 4·63 | flɪdʒ | 3·79 | kɹɪlg | 2·58 |
| tɹɪsk | 5·21 | tɛʃ | 4·63 | blɪg | 3·53 | θwiːks | 2·53 |
| ɹæsk | 5·21 | tiːp | 4·63 | zeɪps | 3·47 | smiːlθ | 2·47 |
| spæk | 5·16 | baɪz | 4·58 | ʧuːl | 3·42 | smɛɹf | 2·47 |
| gɛɹ | 5·11 | glɪp | 4·53 | ʃaɪnt | 3·42 | ploʊmf | 2·42 |
| ʃɻn | 5·11 | plɪm | 4·37 | gwɛndʒ | 3·32 | dwoʊdʒ | 2·29 |
| taɹk | 5·11 | ʧaɪnd | 4·37 | ʃɹʊks | 3·32 | ploʊnθ | 2·26 |
| deɪp | 5·11 | guːd | 4·32 | nʌŋ | 3·28 | θeɪpt | 2·26 |
| skɛl | 5·11 | bleɪf | 4·21 | skwalk | 3·26 | smiːnθ | 2·06 |
| glɪt | 5·11 | gɛz | 4·21 | twuː | 3·17 | spɹaɹf | 2·05 |
| ʧeɪk | 5·05 | zeɪ | 4·16 | smʌm | 3·05 | pwʌdz | 1·74 |

REFERENCES

Albright, Adam (2002a). Islands of reliability for regular morphology: evidence from Italian. *Lg* **78**. 684–709.

Albright, Adam (2002b). The lexical bases of morphological well-formedness. In Sabrina Bendjaballah, Wolfgang U. Dressler, Oskar E. Pfeiffer & Maria D. Voeikova (eds.) *Morphology 2000*. Amsterdam & Philadelphia: Benjamins. 5–15.

Albright, Adam & Bruce Hayes (2002). Modeling English past tense intuitions with minimal generalization. In Mike Maxwell (ed.) *Proceedings of the 6th Meeting of the ACL Special Interest Group on Computational Phonology*. Philadelphia: Association for Computational Linguistics. 58–69.

Albright, Adam & Bruce Hayes (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* **90**. 119–161.

Auer, Edward T., Jr. & Paul A. Luce (2005). Probabilistic phonotactics in spoken word recognition. In David B. Pisoni & Robert E. Remez (eds.) *The handbook of speech perception*. Malden, Mass. & Oxford: Blackwell. 610–630.

Baayen, R. H. (2008). *Analyzing linguistic data : a practical introduction to statistics using R*. *Cambridge*: Cambridge University Press.

Baayen, R. H., Richard Piepenbrock & Hedderik van Rijn (1993). *The CELEX lexical data base*. [CD-ROM.] Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

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

Bates, David, Martin Maechler & Bin Dai (2008). The lme4 Package (Version 0.999375-28): linear mixed-effects models using S4 classes. Available (March 2009) at http://cran.r-project.org/web/packages/lme4/lme4.pdf.

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

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

Berwick, Robert C. (1986). Learning from positive-only examples: the subset principle and three case studies. In Jaime Guillermo Carbonell, Ryszard S. Michalski & Tom M. Mitchell (eds.) *Machine learning : an artificial intelligence approach*. Vol. 2. Los Altos, Ca.: Kaufmann. 625–645.

Buchwald, Adam (2007). Determining well-formedness in phonology: type vs. token frequency. Paper presented at the 81st Annual Meeting of the Linguistic Society of America, Anaheim.

Bybee, Joan (1995). Regular morphology and the lexicon. *Language and Cognitive Processes* **10**. 425–455.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Clements, G. N. & Samuel J. Keyser (1983). *CV phonology : a generative theory of the syllable*. Cambridge, Mass.: MIT Press.

Coleman, John & Janet Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In John Coleman (ed.) *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, NJ: Association for Computational Linguistics. 49–56.

Coltheart, Max, Eileen Davelaar, Jon Torfi Jonasson & Derek Besner (1977). Access to the internal lexicon. In Stan Dornic (ed.) *Attention and performance*. Vol. 6. Hillsdale, NJ: Erlbaum. 535–555.

Davis, Stuart (1984). Some implications of onset-coda constraints for syllable phonology. *CLS* **20**. 46–51.

Frisch, Stefan A., Nathan R. Large & David 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, Joseph H. & James J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *Word* **20**. 157–177.

Hammond, Michael (1999). *The phonology of English : a prosodic optimality-theoretic approach*. Oxford: Oxford University Press.

Hay, Jennifer (2003). *Causes and consequences of word structure*. New York & London: Routledge.

Hay, Jennifer, Janet Pierrehumbert & Mary E. Beckman (2004). Speech perception, well-formedness and the statistics of the lexicon. In John Local, Richard Ogden & Rosalind Temple (eds.) *Phonetic interpretation : papers in laboratory phonology* VI. Cambridge: Cambridge University Press. 58–74.

Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. In Kager *et al*. (2004). 158–203.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**. 379–440.
Jurafsky, Daniel & James H. Martin (2000). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.
Jusczyk, Peter W., Paul A. Luce & Jan Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* **33**. 630–645.
Kager, René, Joe Pater & Wim Zonneveld (eds.) (2004). *Constraints in phonological acquisition*. Cambridge: Cambridge University Press.
Kendall, Maurice G. (1990). *Rank correlation methods*. 5th edn. New York: Oxford University Press.
Kessler, Brett & Rebecca Treiman (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language* **37**. 295–311.
Luce, Paul A. (1986). Neighborhoods of words in the mental lexicon. Technical report, Speech Research Laboratory, Department of Psychology, Indiana University.
Luce, Paul A. & David B. Pisoni (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing* **19**. 1–36.
Newman, Rochelle S., James R. Sawusch & Paul A. Luce (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology : Human Perception and Performance* **23**. 873–889.
Nosofsky, Robert M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology : General* **115**. 39–57.
Ohala, John J. & Manjari Ohala (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In John J. Ohala & Jeri Jaeger (eds.) *Experimental phonology*. Orlando: Academic Press. 239–252.
Pierrehumbert, Janet B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.) *Frequency and the emergence of linguistic structure*. Amsterdam & Philadelphia: Benjamins. 137–157.
Pinheiro, José C. & Douglas M. Bates (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
Pitt, Mark A. & James M. McQueen (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* **39**. 347–370.
Prasada, Sandeep & Steven Pinker (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* **8**. 1–56.
Prince, Alan & Paul Smolensky (2004). *Optimality Theory : constraint interaction in generative grammar*. Malden, Mass. & Oxford: Blackwell.
Prince, Alan & Bruce Tesar (2004). Learning phonotactic distributions. In Kager *et al.* (2004). 245–291.
R Development Core Team (2008). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available (March 2009) at http://www.r-project.org.
Ramscar, Michael & Daniel Yarlett (2003). Semantic grounding in models of analogy: an environmental approach. *Cognitive Science* **27**. 41–71.
Saul, Lawrence & Fernando Pereira (1997). Aggregate and mixed-order Markov models for statistical language processing. In Claire Cardie & Ralph Weischedel (eds.) *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. Somerset, NJ: Association for Computational Linguistics. 81–89.
Scholes, Robert J. (1966). *Phonotactic grammaticality*. The Hague: Mouton.
Schütze, Carson T. (2005). Thinking about what we are asking speakers to do. In Stephan Kepser & Marga Reis (eds.) *Linguistic evidence : empirical, theoretical and computational perspectives*. Berlin & New York: Mouton de Gruyter. 457–484.
Selkirk, Elizabeth O. (1982). The syllable. In Harry van der Hulst & Norval Smith (eds.) *The structure of phonological representations*. Part 2. Dordrecht: Foris. 337–384.

Sendlmeier, Walter F. (1987). Auditive judgements of word similarity. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **40**. 538–547.
Shademan, Shabnam (2006). Is phonotactic knowledge grammatical knowledge? *WCCFL* **25**. 371–379.
Shademan, Shabnam (2007). *Grammar and analogy in phonotactic well-formedness judgments*. PhD thesis, University of California, Los Angeles.
Sheskin, David J. (2004). *Handbook of parametric and nonparametric statistical procedures*. 3rd edn. Boca Raton: Chapman & Hall.
Tessier, Anne-Michelle (2006). *Biases and stages in phonological acquisition*. PhD dissertation, University of Massachusetts, Amherst.
Vitevitch, Michael S. & Paul A. Luce (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science* **9**. 325–329.
Vitevitch, Michael S. & Paul A. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* **40**. 374–408.
Vitevitch, Michael 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.
Vitevitch, Michael S. & Paul A. Luce (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language* **52**. 193–204.
Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce & David Kemmerer (1996). Phonotactic and metrical influences on adult ratings of spoken nonsense words. *Proceedings of the International Conference on Spoken Language Processing* (*ICSLP*). 82–85.
Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce & David Kemmerer (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech* **40**. 47–62.