# Categorical and gradient aspects of wordlikeness*

Kyle Gorman
University of Pennsylvania

Revised January 2013 (comments welcome)

**Abstract**

Gradient wordlikeness judgements do not necessarily imply that there is a gradient well-formedness system underlying them; gradient judgements may be an artifact of gradient rating tasks. Dubious architectural assumptions are needed for speakers to report gradient well-formedness judgements. Simple baselines better account for gradient well-formedness judgements than state-of-the-art computational models of gradient phonotactic knowledge.

## 1 Introduction

Much of the recent work in phonotactic theory has attempted to incorporate the intuition that phonotactic wellformedness is not an "all-or-nothing" matter. Rather, it is alleged, well-formedness judgements have more fidelity than is implied by a simple contrast between "possible" and "impossible", and therefore must be measured and studied at a greater degree of granularity.

This is hardly a novel claim, though it has taken on greater import with the emergence of computational models of wordlikeness. Early generative discussions of wordlikeness (e.g., Chomsky and Halle 1965, Halle 1962) are best remembered for the famous examples [blɪk] and [bnɪk], the former representing a "possible word" of English and the latter representing an "impossible word". A naïve account of this contrast would be to derive it from the assumption that segments must be parsed into syllables or subject to further phonological repair (e.g., Hooper 1973:10f., Kahn 1976:57f., Itô 1989, Wolf and McCarthy 2009:19f.). Unlike some languages (e.g., Morroccan Arabic: *bniqa* 'closet'), English does not permit stop-nasal onsets like [bn], so the latter nonce word cannot surface as such. In other words, [bnɪk] is an impossible surface representation in English. However, in *The Sound Pattern of English* (henceforth, *SPE*), Chomsky and Halle (1968) introduce a third nonce word, [bznk], constructed so as to be even less English-like than [bnɪk].[1] This leads Chomsky and Halle to conclude that wordlikeness intuitions are gradient.

*Thanks to Gene Buckley, Constantine Lignos, Hilary Prichard, Charles Yang, and audiences at NELS 43 for helpful comments.

[1]That wordlikeness judgements depend on language-specific knowledge is apparent given that [bznk] is not impossible in all languages: Imdlawn Tashlhiyt Berber permits whole words consisting of a stop-fricative-nasal-stop sequence (e.g., [tzmt] 'it is stifling'; Dell and Elmedlaoui 1985:112). This is clear evidence for the assumption that wordlikeness depends on language-specific knowledge.

> Hence, a real solution to the problem of "admissibility" will not simply define a tripartite categorization of occurring, accidental gap, and inadmissible, but will define the 'degree of admissibility' of each potential lexical matrix in such a way as to…make numerous other distinctions of this sort (*SPE*:416–417)

This brings the theory of wordlikeness in line with the view of syntactic grammaticality presented by foundational documents like *The Logical Structure of Linguistic Theory* (Chomsky 1955) and *Aspects of the Theory of Syntax* (Chomsky 1965), and reflexes can be found in later work, such as the proposals of Chomsky (1986) and Huang (1982); see Schütze 1996:43f. for a critique.

Chomsky and Halle's claim about the gradient nature of phonotactic wellformedness does not seem to have had much of an impact on practices of the time—as can be seen from discussion in the previous chapter, contemporary critiques focused on other elements of the *SPE* phonotactic theory—but reflexes can once again be found in later work: for instance, Borowsky (1989), Clements and Keyser (1983:50f.), and Myers (1987) all assume a contrast between "peripheral" and absolutely ungrammatical sound sequences.

Recent discussions of gradient grammaticality in wordlikeness attempt to present experimental support for Chomsky and Halle's intuitions:

> 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. (Albright 2009:9)
>
> In the particular domain of phonotactics gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale. (Hayes and Wilson 2008:382)
>
> A defect of current grammatical accounts of phonotactics is that they render simple up-or-down decisions concerning well-formedness and cannot account for gradient judgements. But when judgements are elicited in a controlled fashion from speakers, they always emerge as gradient, including all intermediate values. (Shademan 2006:371)

If the presence of intermediate values in wordlikeness tasks is evidence for the gradient nature of phonotactic wellformedness, then it follows that wordlikeness intuitions should be measured and modeled with a high degree of granularity. For instance, this would be strong evidence against the naïve account of the [blɪk]-[bnɪk] contrast just alluded to, since it cannot easily be extended to account for "numerous other distinctions". However, this chapter argues that there are both theoretical and empirical reasons to doubt the implicit hypothesis linking scalar wordlikeness ratings and gradient wellformedness. First, intermediate ratings are characteristic of all gradient rating tasks, and therefore are irrelevant to the question of whether the internal phonotactic system is categorical or gradient. Secondly, simple baselines better account for gradient well-formedness judgements than current computational models of phonotactic knowledge, suggesting that the gradience observed in these tasks do not derive from known grammatical mechanisms.

## 2 Aspects of the theory of gradient grammaticality

The aforementioned discussions of gradient aspects of wordlikeness judgements takes for granted that intermediate ratings are the product of an internal system of gradient grammaticality. This view is itself an instance of what is known as *common-sense* or *naïve realism*;

in the cognitive sciences, this often takes the form of the assumption that experimental data can be taken at face value, without mediation from other sources of information. However, there are several arguments for *a priori* skepticism about the (a) linguistic abilities required for reporting gradient grammaticality judgements, (b) intermediate acceptability ratings as evidence for gradient grammaticality, and (c) the total lack of previous attempts to consider categorical models of wellformedness.

## 2.1 A model of gradient intuitions

Current research into gradient wellformedness is concerned with specifying a function from sound sequences to scalar judgements, and thus describes the wellformedness system at a level of some abstraction, corresponding roughly to what Marr (1982) calls the "computational" level of description. This is only one part of any model of gradient grammaticality, however; further assumptions are needed to articulate the internal representations and algorithms by which this function computes.

First, consider the architecture which is implied by any system of gradient grammaticality. It is essential that a system of gradient grammaticality have access to a relatively faithful representation of stimuli in a wellformedness task, and therefore it must be able to parse an enormous range of linguistic structures, including many which are not generated by the grammar itself; independent perception and production grammars may be necessary. A scalar value, representing wellformedness, must then be assigned to this parse. To report wellformedness, the speaker must transform this scalar value in accordance with the numerical scale chosen by the experimenter.

Each step of this procedure merits scrutiny, however. First, speakers have difficulty perceiving (Berent et al. 2007, Brown and Hildum 1956, Dupoux et al. 1999, Kabak and Idsardi 2007) and producing (Davidson 2005, 2006a,b, 2010, Gallagher in press, Rose and King 2007, Vitevitch and Luce 1998, 2005) phonotactically illicit non-words, suggesting that speakers' ability to faithfully parse illicit representations is at best quite limited. Secondly, the computation of a scalar value serves no further purpose than to provide for gradient well-formedness judgements, so an objection might be made here on evolutionary grounds. It is quite mysterious why the human language endowment includes constraints on pronominal binding, for instance, but no one can doubt that these constraints are implicated in everyday language use; far more bizarre is the suggestion that the linguistic endowment includes mechanisms only implicated in certain experimental tasks. Next, speakers must be able to consciously access and report the magnitude of this value (it must be *cognitively penetrable* in the sense of Pylyshyn 1984), an ability which is limited in many other domains. Finally, there is some evidence that speakers do not (or perhaps, cannot) respect the numerical scales chosen by experimenters (Sprouse 2011).

It is informative to compare this baroque model to the architecture implied by a binary well-formedness judgement. When presented with a linguistic item in a judgement task, the grammar attempts to assign a parse. Speakers then access whether or not parsing was successful. There are reasons to think that parsing of ungrammatical structures does in fact result in a "crash": whereas syntactic priming increases the acceptability of grammatical structures (Luka and Barsalou 2005), ungrammatical structures show no priming effects (Sprouse 2007). As priming of linguistic structures is thought to implicate shared representations in memory, this suggests different memory mechanisms for grammatical and ungrammatical linguistic objects. Finally, the fact that requests for repetition and clarification are ubiquituous in spontaneous speech illustrates further that speakers are frequently aware when parsing has failed.

Consequently, a great deal of evidence is needed to reject this simple model in favor of the gradient grammaticality architecture.

## 2.2 What some linguistic intuitions might not be

As first noted by Chomsky and Miller (1963), speakers experience difficulty processing sentences with multiple center embeddings. Gibson and Thomas (1999) perform a controlled experiment which reveals that speakers rate sentences like (1a), which is well-formed, less grammatical than (1b), despite the fact that a moment of reflection reveals that the latter sentence is nonsensical.

(1) A well-formedness illusion:

a. The patient who the nurse who the clinic had hired admitted met Jack.
b. *The patient who the nurse who the clinic had hired met Jack.

It is informative to consider that this well-known result has had no effect on the theory of syntactic representations, only on the theory of linguistic memory; it is recognized as the product of cognitive restrictions found in non-linguistic domains, as a *task effect*. This contrasts with the argument made by Hayes (2000), that gradient wordlikeness judgements demand a considerable and unmotivated revision to the grammatical architecture, as is discussed below.

The results of controlled experiments are often biased by subtle details that seem orthogonal to the task: for instance, certain types of duration judgements are systematically biased by consumption of caffeine (Gruber and Block 2005). It should come as no surprise, then, that a highly salient aspect of a judgement task, the scale used for responses, also influences the results obtained. Armstrong, Gleitman, and Gleitman (1983) argue that rating tasks using many-valued scales may induce intermediate ratings as a task effect.

Armstrong et al. (1983) are concerned with experimental evidence for the nature of cognitive concepts. While they do not attempt to dispute that certain concepts (e.g., *fruit*) have a family-resemblance structure (e.g., Rosch 1975), they assert that it is apparent that other concepts are "definitional" (i.e., all-or-nothing), a notion which they illustrate with *odd number*.

> No integer seems to sit on the fence, undecided as to whether it is quite even, or perhaps a bit odd. No odd number seems odder than any other odd number. (Armstrong et al. 1983:274)

However, when subjects are asked to rate, using a 7-point Likert scale, how representative individual odd counting numbers are of the concept *odd number*, they freely use intermediate ratings; the ratings they obtain with instances of *odd number* and *even number* are shown in Figure 1.

This suggests that the gradience observed is primarily an artifact of the task itself. Schütze suggests that the nature of this effect might be understood as the result of speakers' attempts to reconcile bizarre experimental tasks with their knowledge.

> Putting it another way, when asked for gradient responses, participants will find some way to oblige the experimenter; if doing so is incompatible with the experimenter's actual question, they apparently infer that she must have really intended to ask something slightly different. (Schütze 2011:24)

As Armstrong et al. observe, these results show that the scalar judgement tasks provide no evidence as to whether the category being rated is categorical or gradient.

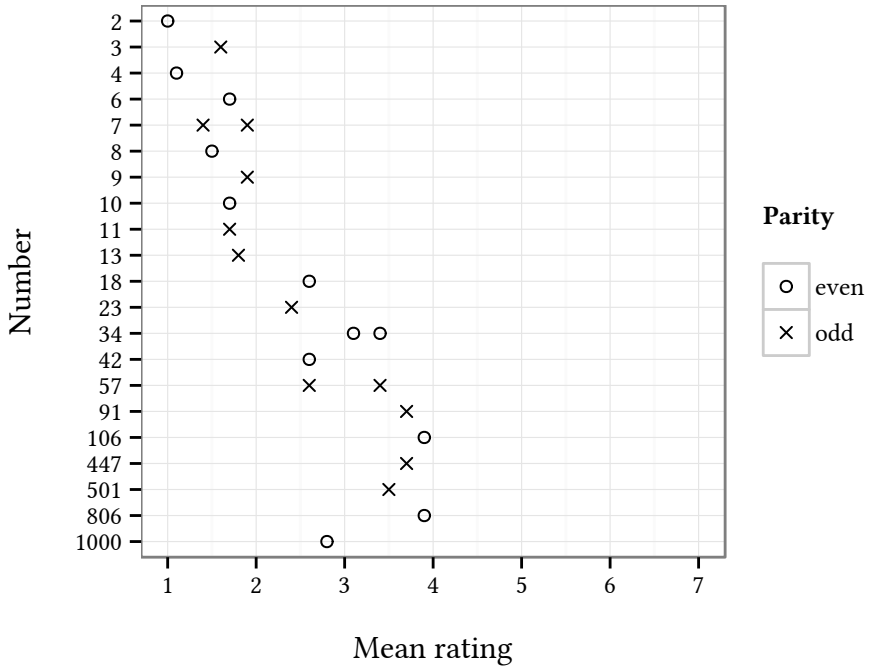


Figure 1: When asked to rate how representative even and odd numbers were of "even" and "odd", respectively, subjects use intermediate ratings (from Armstrong et al. 1983; "1" indicates "most representative")

> …we hold that *fruit* and *odd number* have different structures, and yet we obtain the same experimental outcome for both. But if the same result is achieved regardless of the concept structure, then the experimental design is not pertinent to the determination of concept structure. (Armstrong et al. 1983:284–5)

It might be said that these results reveal something about the representation of odd numbers. Armstrong et al. anticipate this objection.

> Some have responded to these findings very consistently, by asserting that the experimental findings are to be interpreted as before: that, psychologically speaking, odd numbers as well as birds and vegetables are graded concepts… We reject this conclusion just because we could not explain how a person could compute with integers who believed that 7 was odder than 23. We assert confidently that the facts about subjects being able to compute and about their being able to give the definition of odd number, etc., are the more important, highly entrenched, facts we want to preserve and explain… (Armstrong et al. 1983:284)

No scientist has risen to the challenge of constructing a theory that might account for the fact that 447 is rated more odd than 3, and as Armstrong et al. suggest, it is unclear whether such a theory could preserve more central facts about oddness. Here it is possible to draw an analogy to phonotactic theory. According to current orthodoxy, the wellformedness of a sequence is closely related to its type frequency (i.e., frequency in the lexicon). Is it then the case that [bl] is a significantly "better" onset than [kl], simply because the former is

approximately twice as frequent, and if so, is it possible to also ensure that these sequences are treated the same with respect to syllabification?

## 2.3 Evidencing gradience

Hayes (2000) argues that it is "uninsightful" to attribute gradience to task effects, insofar as these effects implicate grammatical representations.

> ...patterns of gradient well-formedness often seem to be driven by the very same principles that govern absolute well-formedness... I conclude that the proposed attribution of gradient well-formedness judgments to performance mechanisms would be uninsightful. Whatever "performance" mechanisms we adopted would look startlingly like the grammatical mechanisms that account for non-gradient judgments. (Hayes 2000:99)

The logic of this implication is indisputable. However, there is little empirical support for the claims that absolute and gradient well-formedness derive from similar principles; indeed, there have been no prior attempts to evaluate categorical and gradient models of wordlikeness on an equal footing. In light of the complexities of gradient models, such an evaluation requires strong quantitative evidence for the superiority of gradient grammatical models. The evaluation below represents a first attempt to fill this gap.

It is not that categorical models have been ignored by the literature on wordlikeness modeling, but rather that they have not been compared. Frisch et al. (2000) and Vitevitch et al. (1997) find that speakers' wordlikeness ratings of multisyllabic words are correlated with a probabilistic measure of the well-formedness of the constituent syllables. Unfortunately, no attempt is made to control for the well-formedness of syllable contact clusters in these words: some of the stimuli have medial consonant clusters containing both voiced and voiceless obstruents (e.g., [gaɪbsaɪk]), something which is exceptionally rare in English simplex words (Gorman in press). Similarly, Hayes and Wilson (2008), who compare their gradient model of wordlikeness against a set of English phonotactic constraints proposed by Clements and Keyser (1983), first transform these constraints, several of which are without exception, into probabilities. While this is consistent with their claim, that "the ability to model gradient intuitions [is] an important criterion for evaluating phonotactic models" (Hayes and Wilson 2008:382), little insight can be gained by annotating an exceptionless rule "$p = 1.0$". Hayes and Wilson's principle precludes any attempt to test the hypotheses that underlies it, and therefore must be rejected.

# 3 Evaluation

As Newmeyer (2007) writes "the idea that categoricity is not represented in the data itself is a truism. Whether distinctions of grammaticality (as opposed to acceptability) are binary is a difficult question." (398) The mere presence of gradience in judgements cannot falsify the claim of gradient grammaticality; another method is needed to evaluate this claim. As a first step towards a falsifiable theory of gradient wordlikeness, the remainder of this chapter considers intermediate ratings in gradient wordlikeness tasks are reliably predicted by the computational models that have been proposed. If a model is incapable of accounting for intermediate ratings, there are two possibilities: either the model itself is improperly specified

| | subjects | items | trials |
|---|---|---|---|
| Albright | 68 | 40 | 2,720 |
| Albright and Hayes | 24 | 86 | 2,064 |
| Scholes | 33 | 63 | 2,178 |
| Total | 125 | 187 | 6,962 |

Table 1: Subject and item counts for the wordlikeness study

and therefore at fault, or the inputs and outputs of the model are unrelated to the actual causes of the intermediate ratings, *contra* Hayes (2000).

It is plausible that speakers might differentiate, in a regular fashion, between different types of "impossible" words, and a gradient model should reliably predict the distinctions that speakers make. There are also claims that speakers distinguish between different types of "possible" words, so that, for instance, [stɪn] *stin* is rated more English-like than [blɪn] *blin* (e.g., Albright 2009), because the former onset is more frequent in the English lexicon. Even if wordlikeness judgements can be effectively modeled with a gross contrast between possible and impossible words, a gradient model might show a correlation with the residual ratings. All of these possibilities are considered below.

## 3.1 Materials

This evaluation uses a large sample of three previously published studies on English wordlikeness comprising 125 subjects and 187 items. Two criteria were used to select these three studies. First, the stimuli must be presented aurally so as to eliminate any possibility of orthographic effects (e.g., Berent et al. 2001, Berent 2008). Secondly, the data must be sufficiently "phonotactically diverse": that is, it must include both items like *blick* and *bnick*. This excludes studies like that of Bailey and Hahn (2001), in which few if any items contain gross phonotactic violations of the type represented by *bnick*. In the absence of phonotactic violences, little variance in wordlikeness ratings can be attributed to phonotactic wellformedness, making it difficult to determine the coverage of gradient wellformedness models. The data used here is summarized in Table 1.

### 3.1.1 Albright 2007

Albright (2007) administers a wordlikeness task in which 68 adult speakers rate 40 monosyllabic nonce words, presented aurally, on a 7-point Likert scale with endpoints labeled "completely impossible as an English word" and "would make a fine English word". Albright's study is primarily concerned with the effects of different onset types (e.g., well-formed /bl/, marginal /bw/, unattested /bn, bd, bz/), and there is less diversity among the choice of rimes, none of which are obviously ill-formed.

### 3.1.2 Albright and Hayes 2003 (norming experiment)

Albright and Hayes (2003) have 24 adult speakers rate 87 aurally presented monosyllabic nonce words on a 7-point Likert scale with endpoints labeled "completely bizarre, impossible as an English word" and "completely normal, would make a fine English word". This task was

administered to establish phonotactic norms for a later nonce word inflection task. Their item [raɪf] is excluded in this study, since this is an actual word of English, *rife*. Albright (2009) uses this data to compare computational models of wordlikeness.

### 3.1.3 Scholes 1966 (experiment 5)

Scholes (1966) conducts several wordlikeness tasks with students in 7th grade (approximately 12–13 years of age). The data used here is his experiment 5, in which 33 speakers provide a "yes" or "no" as to whether each of the 63 items, presented aurally, are "likely to be usable as a word of English". Like the study by Albright (2007), the focus is on onset well-formedness and there is minimal diversity in rime type. Two items, [klʌŋ] *clung* and [bɹʌŋ] *brung* (a dialectal past participle of *bring*), are excluded here as actual words of English. Albright (2009) and Hayes and Wilson (2008) also use this data for the purposes of model evaluation; following Frisch et al. (2000), they use the proportion of "yes" responses for each item so as to derive a continuous measure of well-formedness.

## 3.2 Method

Models are evaluated by comparing their scores to the average rating of each word using four correlation statistics. Each of these range between $[-1, 1]$, where 1 indicates a perfect positive correlation and $-1$ denotes a perfect negative correlation. Hayes and Wilson (2008) evaluate their model using the Pearson ("product-moment") $r$, a parametric correlation measure. It has long been argued (e.g., Stevens 1946) that parametric statistics are inappropriate for analysis of Likert scale data, like those used by Albright (2007) and Albright and Hayes (2003), since the Pearson $r$ makes a *linearity assumption*. That is, it assumes that nonce words rated "1" and "3", for instance, are just as different as are those rated "4" and "6". A weaker assumption, more appropriate for Likert scale data, is the *monotonicity assumption*: that "1" is less English-like than "3", which is less English-like than "4", and so on. However, it also has been claimed that $r$ is particularly robust to violations of the linearity assumption (e.g., Havlicek and Peterson 1976). Pearson $r$ is reported here, but this should not be taken to imply an endorsement of its use for Likert scale data.

Hayes and Wilson also report Spearman $\rho$; this statistic requires only the weaker assumption of monotonicity, but it is difficult to give a simple interpretation to the coefficient. Two other non-parametric statistics, the Goodman-Kruskal $\gamma$ and the Kendall $\tau_b$ are much easier to interpret, as follows (Noether 1981). These statistics are computed by comparing every model score/wordlikeness rating pair to every other: a comparison is counted as *concordant* if the greater of the two model scores is the one associated with the greater of the two wordlikeness ratings (that is, the model ranks these two nonce words in accordance with speakers' ratings), and as *discordant* otherwise. These two statistics differ only in the treatment of "ties", pairs where either the model score or wordlikeness rating are identical. For $\gamma$, ties are ignored, and the coefficient is

$$\gamma = \frac{c - d}{c + d}$$

where $c$ and $d$ represent the number of concordant and discordant pairs, respectively. The $\tau_b$ statistic uses a similar formula, but also incorporates a penalty for ties in model score which are not also paired with ties in wordlikeness ratings, or vice versa. Albright (2009) uses a variant of this statistic to evaluate wordlikeness models.

## 3.3 Models

The nonce word stimuli from these three studies are scored automatically using four computational models. The first two models represent baselines for comparison to the latter two state-of-the-art gradient models. The scores are reproduced in Appendix A.

### 3.3.1 Gross phonotactic violation

A simple baseline is constructed by separating nonce words into those which contain a phonotactic violation and those which do not. As all nonce words here are monosyllabic, this task can be localized to two subcomponents of the syllable, the onset and the rime. This is not to imply that these are the only domains over which phonotactic violations might be stated, but there are prior claims that onset and rime are particularly important domains for stating phonotactic constraints (e.g., Fudge 1969, Kessler and Treiman 1997, Treiman et al. 2000). Speakers are adept at separating syllables into these units (Treiman 1983, 1986, Treiman et al. 1995), and they are implicated by patterns of speech errors (Fowler 1987, Fowler et al. 1993).

Operationalizing "phonotactic violation" is somewhat more difficult. The simplest possible mechanism is chosen here: an onset or rime is identified as well-formed if it occurs with non-zero frequency in a representative sample, and is identified as ill-formed otherwise. This is not to imply that all unattested onsets or rimes should be regarded as ill-formed, or that all onsets or rimes with non-zero frequency in this data are well-formed. For instance, Albright (2009) judges [dɹɛsp] *dresp* to be phonotactically well-formed, despite the total lack of [ɛsp] rimes in English; similar observations have been made concerning English onsets (e.g., Cairns 1972, Moreton 2002).

The representative sample used to define the phonotactic baseline is derived from those entries of the CMU pronunciation dictionary which occur at least once per million words in the SUBTLEX-US frequency norms; these norms are known to be particularly strongly correlated with other behavioral measures (Brysbaert and New 2009). These pronunciations are then syllabified, and individual syllables parsed into onset and rime, according to a process described in detail in Appendix B.

Wordlikeness ratings from the three studies are plotted against this gross contrast in Figure 2. While there are several outliers, there can be little doubt that gross phonotactic status accounts for a considerable amount of variance in wordlikeness judgements.

### 3.3.2 Lexical neighborhood density

A second baseline is provided by measures of similarity to existing English words, which has long been applied to model wordlikeness judgements (e.g., Bailey and Hahn 2001, Greenberg and Jenkins 1964, Kirby and Yu 2007, Ohala and Ohala 1986, Shademan 2006, 2007, Vitevitch and Luce 1998, 1999). Chomsky (1955: 151, fn. 27) suggests that grammaticality judgements in general might be influenced by similarity to existing grammatical structures, and Chomsky and Halle (1968:417f.) outline a similarity-based wordlikeness model. More recently, it has been observed (e.g., Coleman and Pierrehumbert 1997:51, Hay et al. 2004) that nonce words which flagrantly violate English sonority restrictions but which bear common affixes (e.g., **mrupation*) are rated highly English-like.

A wide variety of lexical similarity measures were considered, including a variant of the Generalized Neighborhood model (Bailey and Hahn 2001), PLD20 (Suárez et al. 2011), and a set of measures provided by the Irvine Phonotactic Calculator (Vaden et al. 2009). The measure best correlated with wellformedness judgements is also the most venerable measure

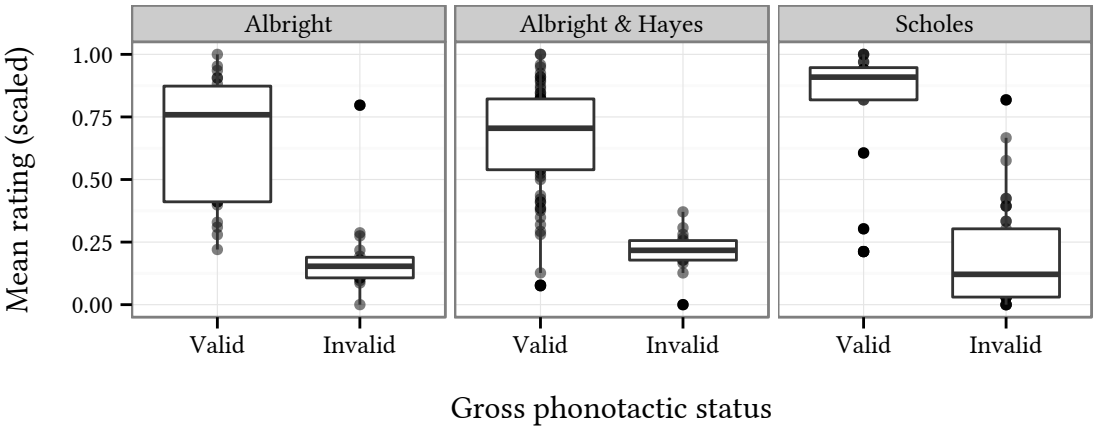


Figure 2: Gross phonotactic status and item-averaged wordlikeness ratings

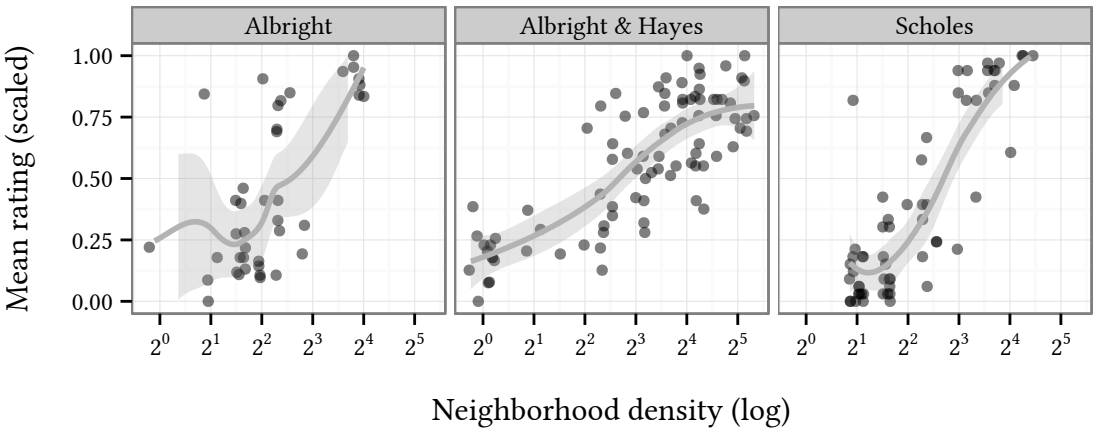


Figure 3: Correlation between Coltheart's *N* and item-averaged wordlikeness ratings, with LOESS curve

of lexical similarity: Coltheart's *N* (Coltheart et al. 1977), which is defined as the number of words in some representative sample which can be changed into a target nonce word by a single insertion, deletion, or substitution of a phoneme. Greenberg and Jenkins (1964) find a correlation between wordlikeness ratings and a variant of this measure which only counts words differing by a single substitution. This is plotted against ratings from the three studies in Figure 3 with a superimposed local regression (LOESS; Cleveland and Devlin 1988) curve; neighborhood density accounts for much of the variance in ratings.

While there is nothing inherently "phonotactic" about Coltheart's *N*, it indirectly incorporates much of the information present in the gross phonotactic baseline. Consider [blɪk]: since there is nothing marked about any part of this nonce word, a "neighbor" might be found by modifying any phone: e.g., *click*, *brick*, *bloke*, *bliss*. However, since [bn] onsets are unattested in English, a neighbor of [bnɪk] must somehow modify this cluster: this leaves only *brick* and *nick*. Bailey and Hahn (2001) and Frauenfelder et al. (1993) note that neighborhood density is also strongly correlated with measures like bigram probability, but it has been ar-

| | Pearson $r$ | Spearman $\rho$ | G-K $\gamma$ | Kendall $\tau_b$ |
|---|---|---|---|---|
| featural bigrams | .71 | .64 | .45 | .45 |
| segmental bigrams | .74 | .67 | .48 | .47 |
| segmental bigrams with smoothing | .75 | .70 | .50 | .50 |

Table 2: Correlation between item-averaged wordlikeness ratings for the Albright and Hayes (2003) norming study and three variants of bigram probability

gued elsewhere that phonotactic measures and neighborhood density have distinct effects (e.g., Berent and Shimron 2003, Pitt and McQueen 1998, Vitevitch and Luce 1998, 1999).

### 3.3.3 Segmental bigram probability

Faciliatory effects of bigram probabilities (i.e., shorter latencies) are reported for other nonce word tasks conducted with adults, including single-word shadowing (Vitevitch et al. 1997, Vitevitch and Luce 1998), same/different judgements (Lipinski and Gupta 2005, Luce and Large 2001, Vitevitch and Luce 1999, 2005), and lexical decision (Pylkkänen et al. 2002). Albright (2009) applies bigram probability as a model of wordlikeness judgements. The bigram probability of a sequence *ijk*, for instance, is defined as

$$\hat{p}(ijk) = p(i|\textsc{start}) \cdot p(j|i) \cdot p(k|j) \cdot p(\textsc{stop}|k)$$

That is, it is the product of sequence-initial *i*, the probability of *j* following *i*, the probability of *k* following *j*, and the probability of the sequence ending after *k*.

Albright (2009) compares two variants of this model, the first operating over segments, the second over sets of features. Unfortunately, the latter model is not described in sufficient detail to allow it to be implemented directly, and there is no publicly available implementation. However, Albright's evaluation, which includes the Scholes (1966) and Albright and Hayes (2003) data, finds an advantage for the segment-based model. In implementing this model, it was found that a slight improvement could be made by preventing any phoneme-to-phoneme transition from having zero probability. This is accomplished by adding 1 to the count of every transition, a technique used in natural language processing and known as Laplace, or "add one", smoothing. As can be seen in Table 2 this results in a slight increase in the correlation between the scores from this model and wellformedness ratings This smoothed segmental bigram score is adopted below. In Figure 4, it is plotted against wordlikeness ratings from three studies.

### 3.3.4 Maximum entropy phonotactics

Hayes and Wilson (2008) present a model which uses the principle of maximum entropy to weigh a large number of competing phonotactic constraints (e.g., Goldwater and Johnson 2003, Jäger 2007). Hayes and Wilson use a complex method to evaluate their model. First, they extract onset sequences from the CMU pronunciation dictionary, and use these to train the model. The model is then used to score the onsets of the Scholes (1966) nonce words. Then they compute a parameter for transforming their model scores so as to maximize the correlation between these transformed scores and wordlikeness ratings, then report the re-

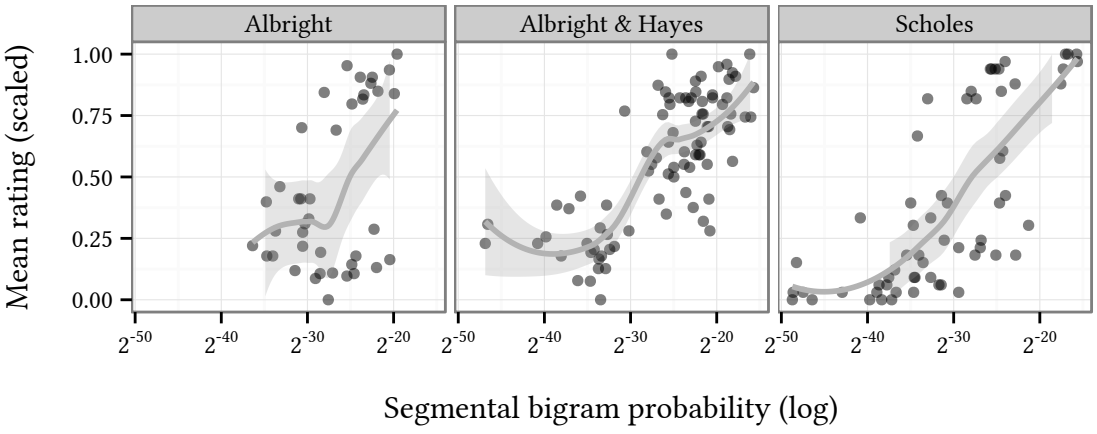


Figure 4: Correlation between smoothed segmental bigram score and item-averaged wordlikeness ratings, with LOESS curve

sulting correlation.[2] Albright (2009) reports that the maximum entropy model, training and testing only on onsets, performs well on the Scholes (1966) data, but does not generalize well to the Albright and Hayes (2003) sample. Consequently, the model was trained to score whole words, not just onsets, using the subset of the CMU dictionary described above.

Since this model has numerous experimenter-defined parameters, a close replication of Hayes and Wilson's original paper is attempted: both their implementation and phonological feature specifications are used here. Following Hayes and White (in press), dictionary entries are syllabified using the procedure described in Appendix B, and a novel feature [±Coda] is added to allow the model to distinguish coda and onset consonants. Also, following Hayes and Wilson, constraints are limited to those spanning as many as three segments and an "accuracy schedule" of [.001, .01, .1, .2, .3] is used. Since the maximum entropy model produces slightly different scores on each run, the worst-performing of 10 runs is reported here, following Hayes and Wilson. The resulting scores are plotted against wordlikeness ratings in Figure 5; it can be seen that the model assigns the highest possible score to a large variety of nonce words, though many words with a low rating receive the highest MaxEnt probability score. It appears that this model is still not robust enough to reliably extracting phonotactic generalizations from monosyllabic words.

## 3.4 Results

Table 3 displays the full set of correlation coefficients, for each of the three data sets, and for each of the four models. The first observation is that in general, there is a positive correlation between model score and ratings in each pair. The two baselines, gross phonotactic status and neighborhood density, are by far the strongest models across statistics and studies, with gross phonotactic status performing the strongest under the Goodman-Kruskal $\gamma$ and on the

[2]This is contrary to standard practices in natural language processing, in that the data used for evaluation is also used to fit the model (namely, the transformation's parameter); when this is the case, there is reason to suspect the parameter values will not generalize to new data. No transformation is used here; this only has an effect on the Pearson $r$ coefficient, since they use a transformation that preserves monotonicity.

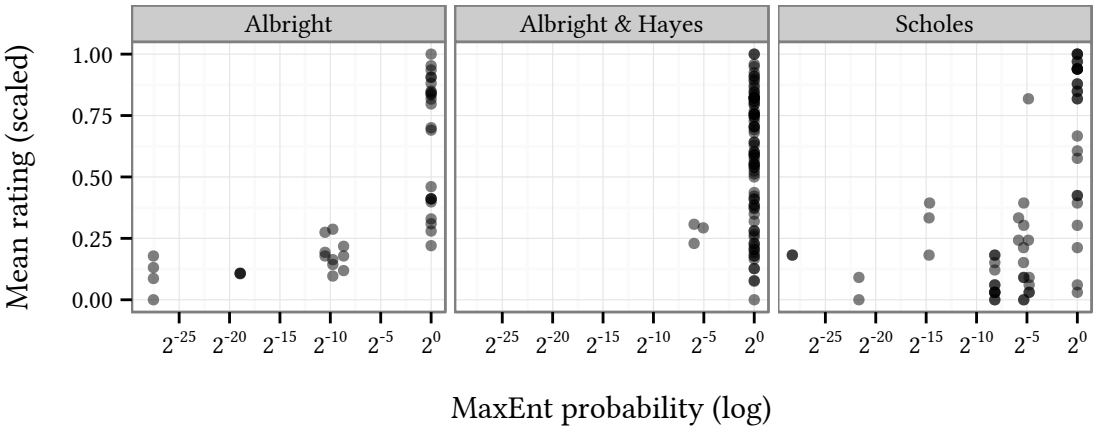


Figure 5: Correlation between MaxEnt score and item-averaged wordlikeness ratings

| | Pearson $r$ | | | Spearman $\rho$ | | | G-K $\gamma$ | | | Kendall $\tau_b$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | AH | S | A | AH | S | A | AH | S | A | AH | S |
| Gross status | .73 | .60 | .80 | .82 | .66 | .80 | .87 | .93 | .91 | .67 | .47 | .62 |
| Density (*N*) | .67 | .79 | .86 | .61 | .74 | .82 | .49 | .57 | .74 | .45 | .56 | .67 |
| Bigram $p$ | .46 | .75 | .74 | .34 | .70 | .79 | .25 | .50 | .63 | .25 | .50 | .61 |
| MaxEnt $p$ | .70 | .21 | .53 | .66 | .39 | .58 | .85 | .61 | .56 | .68 | .16 | .48 |

Table 3: Correlation between item-averaged wordlikeness ratings and model scores

Albright (2007) data, and neighborhood density performing strongly under nearly all other statistics and data sets.

It is also possible to consider whether there is any residual correlation between bigram and MaxEnt model scores, and wordlikeness ratings within the "valid" and "invalid" groups defined by the gross phonotactic status measure. Kendall $\tau_b$ correlations within these subgroups for each data set are shown in Table 4. The only reliable positive correlation is present among the "valid" items as rated by the smoothed segmental bigram model. This model is somewhat capable of accounting for contrasts between different "possible" nonce words: for instance, it favors [plin] *plean* over [brɛlθ] *brelth* just as subjects in the Albright (2007) study do; this can also be seen in the top three panels of Figure 6. Within the set of "invalid" items, however, neither grammatical model reliably distinguishes among items; both models, for instance, rate [ptʌs] *ptus* more well-formed than [bnʌs] *bnus*, but speakers have the opposite preference.

## 3.5 Discussion

The bigram and MaxEnt models do not reliably outperform simple baselines. From this it can be inferred that the gradient models do not reliably predict intermediate ratings. Nor do these models reliably distinguish within classes of "valid" and "invalid" words in a way that conforms with wordlikeness ratings.

| | "Valid" items | | | "Invalid" items | | |
|---|---|---|---|---|---|---|
| | A | AH | S | A | AH | S |
| Bigram score | .65 | .34 | .60 | .03 | −.17 | .47 |
| MaxEnt score | .00 | −.15 | −.32 | −.42 | .29 | −.16 |

Table 4: Kendall $\tau_b$ correlation between model scores and item-averaged wordlikeness ratings, sorted according to gross phonotactic status

A serious limitation of this evaluation is the primitive nature of the gross phonotactic status baseline. It does not allow for any way to state constraints on onset-nucleus sequences, which have been proposed for some languages (e.g., Kirby and Yu 2007 on Cantonese), or constraints spanning whole syllables (e.g., Berkley 1994a,b, Clements and Keyser 1983, Coetzee 2008, Fudge 1969).[3] Furthermore, the gross phonotactic baseline does not have any mechanism for generalizing the wellformedness of [ɛsp] rimes from *clasp*, *lisp*, and other rimes consisting of a lax vowel followed by [sp] found in English, but Borowsky (1989), for instance, proposes a theory of possible rimes in English which makes the correct prediction regarding [ɛsp]. This is not embedded in an acquisitional model, but many models of syllable type acquisition have been proposed (e.g., Fikkert 1994, Levelt et al. 2000, Pan and Snyder 2003, 2004). As observed by Smith (1973) in a careful study of a single child acquiring English, children's productions are at first highly restricted but progress systematically to stages with fewer and fewer restrictions. Assuming productive competence is an appropriate measure of syllable acqusition, this suggests that syllable types are acquired like many other linguistic phenomena in that the child progresses from subset to superset. The difficulty here is that the typology of syllables must be delineated so that, for instance, the robust presence of [æsp] and [ɪsp] implies acceptance of [ɛsp].

The gross phonotactic baseline could also be extended so as to recognize more than two levels of wellformedness, without introducing the infinite amount of contrast implied by fully gradient models. While the bigram and MaxEnt models do not appear to be able to reliably distinguish intermediate levels of well-formedness, it might be desirable to encode the intuition that, for example, [ʒlɪk] *zhlick*, is more English-like than [bnɪk], though both have unattested onsets (e.g., Clements and Keyser 1983:50f.). It is also possible to imagine that phonotactic violations would have a cumulative effect on well-formedness. For instance, a nonce word with an unattested onset and an unattested rime, like [tsɪlm], might be less English-like than either [tsɪl] or [sɪlm], an ability that could easily be extended to the gross phonotactic baseline. Cumulativity effects are predicted by the bigram and MaxEnt models, among others (e.g., Albright et al. 2008, Anttila 1997), but could easily be incorporated into a simple baseline by counting the number of violations. However, as of yet there is no convincing evidence for cumulativity effects in wordlikeness tasks, and the stimuli used here are not suited to test this hypothesis.

[3] It is disputed whether English in particular exhibits onset-nucleus restrictions. Clements and Keyser (1983) claim that "cooccurrence restrictions holding between the nucleus and preceding elements of the syllable appear to be just as common as cooccurrence restrictions holding between the nucleus and following elements" (20), but admit that at least some of these generalizations may represent accidental gaps. However, Kessler and Treiman (1997), argue there are no clear restrictions on English onset-nuclei pairs.

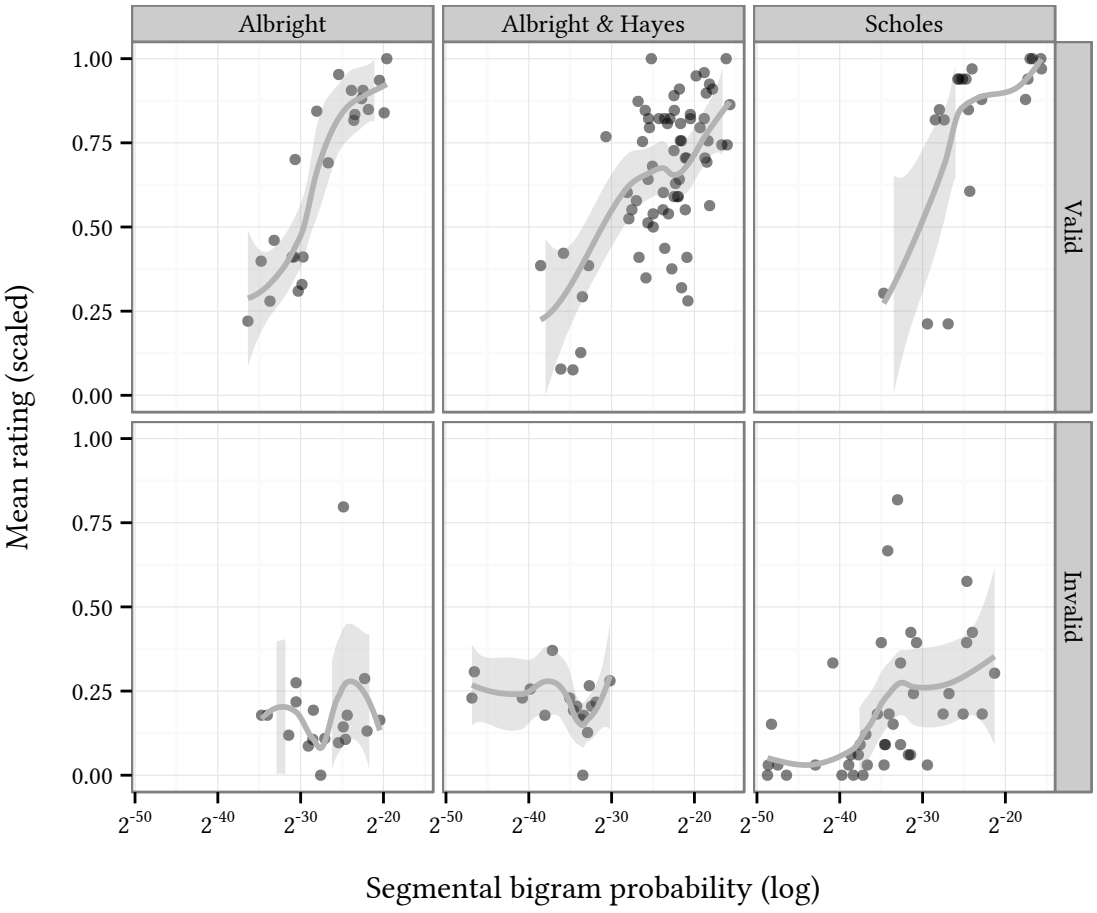


Figure 6: Correlation between smoothed segmental bigram score and item-averaged wordlikeness ratings, sorted according to the gross phonotactic status, with LOESS curve

# 4 Conclusions

State-of-the-art computational models of wellformedness do not reliably predict intermediate ratings in wordlikeness tasks. To the degree to which the bigram or MaxEnt models are correlated with speakers' judgements, these judgements are more precisely modeled by similarity to existing words, or by a gross contrast between attested and unattested onsets and rimes. While it remains an open question whether future gradient models will account for intermediate judgements, the current evidence suggests that gradient grammaticality is not crucial for modeling gradient wordlikeness judgements. This does not imply that wordlikeness judgements collected using Likert scales or magnitude estimation are tainted; Sprouse and Almeida (submitted) argue that gradient wellformedness measures are better able to detect syntactic violations thought to be categorical than are binary judgements, and it seems likely this result would also hold for wordlikeness tasks. However, intermediate ratings can no longer be taken at face value.

# A English wordlikeness ratings

## A.1 Albright (2007)

| | lexical density | −log $p$ (bigram) | −log $p$ (MaxEnt) | gross status | rating (7-point) |
|---|---|---|---|---|---|
| P L IY1 N | 13 | 13.585 | 0.000 | valid | 5.32 |
| B L AA1 D | 13 | 17.609 | 0.000 | valid | 5.13 |
| P L IY1 K | 11 | 14.200 | 0.000 | valid | 5.06 |
| P L EY1 K | 14 | 15.576 | 0.000 | valid | 4.94 |
| P R AH1 N JH | 3 | 16.546 | 0.000 | valid | 4.94 |
| B L UW1 T | 14 | 15.692 | 0.000 | valid | 4.84 |
| P L IH1 M | 5 | 15.126 | 0.000 | valid | 4.71 |
| B L EH1 M P | 1 | 19.447 | 0.000 | valid | 4.69 |
| B L AH1 S | 14 | 13.806 | 0.000 | valid | 4.67 |
| B L AE1 D | 15 | 16.259 | 0.000 | valid | 4.65 |
| B L IH1 G | 4 | 16.347 | 0.000 | valid | 4.58 |
| P R EH1 S P | 4 | 17.214 | 0.000 | invalid | 4.50 |
| B R EH1 N TH | 4 | 21.255 | 0.000 | valid | 4.11 |
| P R AH1 P T | 4 | 18.487 | 0.000 | valid | 4.07 |
| B R EH1 L TH | 2 | 23.014 | 0.000 | valid | 3.14 |
| P W IH1 S T | 4 | 21.499 | 0.000 | valid | 2.94 |
| B W AH1 D | 2 | 20.596 | 0.000 | valid | 2.94 |
| B W AA1 D | 3 | 21.329 | 0.000 | valid | 2.94 |
| P W AE1 D | 2 | 24.103 | 0.000 | valid | 2.89 |
| P W AH1 S | 4 | 20.684 | 0.000 | valid | 2.61 |
| P W EH1 T | 6 | 20.998 | 0.000 | valid | 2.53 |
| P T IY1 N | 4 | 15.440 | 6.762 | invalid | 2.44 |
| B W AE1 D | 2 | 23.365 | 0.000 | valid | 2.41 |
| B N IY1 N | 2 | 21.180 | 7.296 | invalid | 2.39 |
| P W AH1 D Z | 0 | 25.210 | 0.000 | valid | 2.17 |
| P N IY1 N | 2 | 21.181 | 6.019 | invalid | 2.16 |
| B N AH1 S | 6 | 19.732 | 7.296 | invalid | 2.06 |
| P N EH1 P | 2 | 23.587 | 6.019 | invalid | 2.00 |
| B N AA1 D | 2 | 24.066 | 7.296 | invalid | 2.00 |
| B Z IY1 N | 1 | 16.896 | 19.097 | invalid | 2.00 |
| P T AH1 S | 3 | 14.169 | 6.762 | invalid | 1.94 |
| P T EH1 P | 3 | 17.228 | 6.762 | invalid | 1.86 |
| B Z AH1 S | 2 | 15.237 | 19.097 | invalid | 1.81 |
| P N IY1 K | 2 | 21.796 | 6.019 | invalid | 1.76 |
| B D IY1 K | 2 | 18.773 | 13.131 | invalid | 1.72 |
| B D UW1 T | 3 | 19.781 | 13.131 | invalid | 1.71 |
| B D AH1 S | 4 | 17.041 | 13.131 | invalid | 1.71 |
| P T AE1 D | 3 | 17.622 | 6.762 | invalid | 1.67 |
| B Z AA1 D | 1 | 20.151 | 19.097 | invalid | 1.63 |
| B Z AY1 K | 1 | 19.118 | 19.097 | invalid | 1.28 |

## A.2 Albright and Hayes (2003), norming study

| | lexical density | −log $p$ (bigram) | −log $p$ (MaxEnt) | gross status | rating (7-point) |
|---|---|---|---|---|---|
| S L EY1 M | 15 | 17.469 | 0.000 | valid | 5.84 |
| W IH1 S | 34 | 11.208 | 0.000 | valid | 5.84 |
| P IH1 N T | 26 | 13.046 | 0.000 | valid | 5.67 |
| P AE1 NG K | 18 | 13.723 | 0.000 | valid | 5.63 |
| S T IH1 P | 18 | 12.599 | 0.000 | valid | 5.53 |
| M IH1 P | 33 | 12.345 | 0.000 | valid | 5.47 |
| S T AY1 R | 11 | 15.118 | 0.000 | valid | 5.47 |
| M ER1 N | 34 | 12.872 | 0.000 | valid | 5.42 |
| P L EY1 K | 14 | 15.576 | 0.000 | valid | 5.39 |
| S N EH1 L | 10 | 18.582 | 0.000 | valid | 5.32 |
| S T IH1 N | 18 | 10.899 | 0.000 | valid | 5.28 |
| R AE1 S K | 11 | 15.544 | 0.000 | valid | 5.21 |
| T R IH1 S K | 5 | 17.980 | 0.000 | valid | 5.21 |
| S P AE1 K | 17 | 14.205 | 0.000 | valid | 5.16 |
| D EY1 P | 22 | 14.193 | 0.000 | valid | 5.11 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | G | EH1 | R | | 25 | 13.044 | 0.000 | valid | 5.11 |
| | G | L | IH1 | T | | 14 | 16.830 | 0.000 | valid | 5.11 |
| | S | K | EH1 | L | | 16 | 16.356 | 0.000 | valid | 5.11 |
| | | SH | ER1 | N | | 23 | 15.913 | 0.000 | valid | 5.11 |
| | | T | AA1 | R | K | 18 | 17.702 | 0.000 | valid | 5.11 |
| | | CH | EY1 | K | | 28 | 15.023 | 0.000 | valid | 5.05 |
| | G | L | IY1 | D | | 14 | 16.118 | 0.000 | valid | 5.05 |
| | G | R | AY1 | N | T | 4 | 17.626 | 0.000 | valid | 5.00 |
| | P | R | IY1 | K | | 11 | 13.396 | 0.000 | valid | 5.00 |
| | | SH | IH1 | L | K | 8 | 21.270 | 0.000 | valid | 4.89 |
| | | D | AY1 | Z | | 39 | 12.730 | 0.000 | valid | 4.84 |
| | | N | EY1 | S | | 23 | 14.952 | 0.000 | valid | 4.84 |
| | | T | AH1 | NG | K | 18 | 15.046 | 0.000 | valid | 4.84 |
| S | K | W | IH1 | L | | 6 | 18.210 | 0.000 | valid | 4.83 |
| | | L | AH1 | M | | 35 | 11.569 | 0.000 | valid | 4.79 |
| | | P | AH1 | M | | 30 | 11.121 | 0.000 | valid | 4.79 |
| S | P | L | IH1 | NG | | 14 | 15.573 | 0.000 | valid | 4.72 |
| | G | R | EH1 | L | | 3 | 14.624 | 0.000 | valid | 4.63 |
| | | T | EH1 | SH | | 12 | 14.517 | 0.000 | valid | 4.63 |
| | | T | IY1 | P | | 32 | 12.980 | 0.000 | valid | 4.63 |
| | | B | AY1 | Z | | 35 | 12.821 | 0.000 | valid | 4.58 |
| | G | L | IH1 | P | | 11 | 17.377 | 0.000 | valid | 4.53 |
| | | CH | AY1 | N | D | 18 | 17.747 | 0.000 | valid | 4.37 |
| | P | L | IH1 | M | | 5 | 15.126 | 0.000 | valid | 4.37 |
| | | G | UW1 | D | | 29 | 15.448 | 0.000 | valid | 4.32 |
| | B | L | EY1 | F | | 6 | 19.485 | 0.000 | valid | 4.21 |
| | | G | EH1 | Z | | 17 | 16.466 | 0.000 | valid | 4.21 |
| | D | R | IH1 | T | | 8 | 15.563 | 0.000 | valid | 4.16 |
| | F | L | IY1 | P | | 10 | 15.292 | 0.000 | valid | 4.16 |
| | | Z | EY1 | | | 23 | 15.208 | 0.000 | valid | 4.16 |
| S | K | R | AY1 | D | | 5 | 18.722 | 0.000 | valid | 4.11 |
| | | K | IH1 | V | | 16 | 12.591 | 0.000 | valid | 4.05 |
| | F | L | EH1 | T | | 17 | 16.490 | 0.000 | valid | 4.00 |
| | | N | OW1 | L | D | 19 | 19.101 | 0.000 | valid | 4.00 |
| | S | K | IH1 | K | | 13 | 14.628 | 0.000 | valid | 4.00 |
| | B | R | EH1 | JH | | 7 | 17.318 | 0.000 | valid | 3.95 |
| | K | W | IY1 | D | | 10 | 16.039 | 0.000 | valid | 3.95 |
| | S | K | OY1 | L | | 9 | 19.350 | 0.000 | valid | 3.89 |
| | D | R | AY1 | S | | 12 | 17.758 | 0.000 | valid | 3.84 |
| | F | L | IH1 | JH | | 8 | 17.312 | 0.000 | valid | 3.79 |
| | B | L | IH1 | G | | 4 | 16.347 | 0.000 | valid | 3.53 |
| | | Z | EY1 | P | S | 7 | 24.825 | 0.000 | valid | 3.47 |
| | | CH | UW1 | L | | 17 | 14.492 | 0.000 | valid | 3.42 |
| | | SH | AY1 | N | T | 8 | 18.503 | 0.000 | valid | 3.42 |
| | SH | R | UH1 | K | S | 5 | 26.733 | 0.000 | valid | 3.32 |
| | G | W | EH1 | N | JH | 0 | 22.722 | 0.000 | valid | 3.32 |
| | | N | AH1 | NG | | 19 | 15.754 | 0.000 | valid | 3.28 |
| S | K | W | AA1 | L | K | 1 | 25.752 | 0.000 | invalid | 3.26 |
| | T | W | UW1 | | | 5 | 17.918 | 0.000 | valid | 3.17 |
| | S | M | AH1 | M | | 8 | 14.940 | 0.000 | valid | 3.05 |
| | S | N | OY1 | K | S | 4 | 32.283 | 4.136 | invalid | 3.00 |
| | S | F | UW1 | N | D | 1 | 23.241 | 3.507 | valid | 2.94 |
| | P | W | IH1 | P | | 4 | 20.928 | 0.000 | valid | 2.89 |
| | | R | AY1 | N | T | 8 | 14.412 | 0.000 | valid | 2.89 |
| S | K | L | UW1 | N | D | 0 | 22.661 | 0.000 | invalid | 2.83 |
| | S | M | IY1 | R | G | 0 | 27.601 | 0.000 | invalid | 2.79 |
| | F | R | IH1 | L | G | 3 | 24.299 | 0.000 | invalid | 2.68 |
| | SH | W | UW1 | JH | | 0 | 28.270 | 0.000 | invalid | 2.68 |
| | TH | R | OY1 | K | S | 0 | 32.485 | 4.136 | invalid | 2.68 |
| | T | R | IH1 | L | B | 4 | 22.097 | 0.000 | invalid | 2.63 |
| | K | R | IH1 | L | G | 1 | 23.719 | 0.000 | invalid | 2.58 |
| | S | M | EH1 | R | G | 0 | 22.473 | 0.000 | invalid | 2.58 |
| | TH | W | IY1 | K | S | 2 | 23.984 | 0.000 | invalid | 2.53 |
| | S | M | EH1 | R | F | 0 | 23.136 | 0.000 | invalid | 2.47 |
| | S | M | IY1 | L | TH | 0 | 26.377 | 0.000 | invalid | 2.47 |
| | P | L | OW1 | M | F | 0 | 23.336 | 0.000 | invalid | 2.42 |
| | P | L | OW1 | N | TH | 0 | 22.805 | 0.000 | invalid | 2.26 |
| | | TH | EY1 | P | T | 4 | 23.380 | 0.000 | valid | 2.26 |
| | S | M | IY1 | N | TH | 0 | 25.043 | 0.000 | valid | 2.06 |
| S | P | R | AA1 | R | F | 0 | 24.031 | 0.000 | valid | 2.05 |
| | P | W | AH1 | JH | | 0 | 23.205 | 0.000 | valid | 1.74 |

## A.3 Scholes (1966), experiment 5

| | | | | lexical density | −log $p$ (bigram) | −log $p$ (MaxEnt) | gross status | rating (binary) |
|---|---|---|---|---|---|---|---|---|
| G | R | AH1 | N | 18 | 11.799 | 0.000 | valid | 33 |
| K | R | AH1 | N | 21 | 11.597 | 0.000 | valid | 33 |
| S | T | IH1 | N | 18 | 10.899 | 0.000 | valid | 33 |
| S | M | AE1 | T | 13 | 16.654 | 0.000 | valid | 32 |
| P | R | AH1 | N | 11 | 10.845 | 0.000 | valid | 32 |
| S | L | ER1 | K | 12 | 17.846 | 0.000 | valid | 31 |
| F | L | ER1 | K | 11 | 17.456 | 0.000 | valid | 31 |
| B | L | AH1 | NG | 8 | 17.156 | 0.000 | valid | 31 |
| D | R | AH1 | NG | 7 | 17.753 | 0.000 | valid | 31 |
| T | R | AH1 | N | 12 | 11.975 | 0.000 | valid | 31 |
| F | R | AH1 | N | 12 | 12.177 | 0.000 | valid | 29 |
| S | P | EY1 | L | 16 | 15.851 | 0.000 | valid | 29 |
| S | N | EH1 | T | 7 | 19.384 | 0.000 | valid | 28 |
| P | L | AH1 | NG | 11 | 16.960 | 0.000 | valid | 28 |
| SH | R | AH1 | K | 8 | 19.734 | 0.000 | valid | 27 |
| G | L | AH1 | NG | 9 | 18.990 | 0.000 | valid | 27 |
| M | R | AH1 | NG | 1 | 22.888 | 3.365 | invalid | 27 |
| SH | L | ER1 | K | 4 | 23.711 | 0.000 | invalid | 22 |
| S | K | IY1 | P | 15 | 16.845 | 0.000 | valid | 20 |
| V | R | AH1 | N | 4 | 17.087 | 0.000 | invalid | 19 |
| S | R | AH1 | N | 9 | 16.626 | 0.000 | invalid | 14 |
| V | L | ER1 | K | 2 | 21.777 | 0.000 | invalid | 14 |
| M | L | AH1 | NG | 4 | 21.300 | 10.164 | invalid | 13 |
| SH | T | IH1 | N | 3 | 17.106 | 0.000 | invalid | 13 |
| F | P | EY1 | L | 4 | 24.250 | 3.685 | invalid | 13 |
| ZH | R | AH1 | N | 4 | 28.305 | 4.042 | invalid | 11 |
| F | SH | IH1 | P | 2 | 22.640 | 10.198 | invalid | 11 |
| SH | N | EH1 | T | 2 | 24.044 | 0.000 | valid | 10 |
| F | T | IH1 | N | 2 | 14.767 | 3.685 | invalid | 10 |
| Z | R | AH1 | N | 5 | 21.556 | 4.042 | invalid | 8 |
| N | R | AH1 | N | 5 | 18.588 | 3.365 | invalid | 8 |
| SH | M | AE1 | T | 1 | 20.389 | 0.000 | valid | 7 |
| S | F | IY1 | D | 7 | 18.656 | 3.701 | valid | 7 |
| Z | L | ER1 | K | 2 | 24.578 | 5.678 | invalid | 6 |
| Z | T | IH1 | N | 1 | 23.600 | 5.678 | invalid | 6 |
| F | S | EH1 | T | 4 | 19.079 | 10.198 | invalid | 6 |
| V | Z | IH1 | P | 1 | 17.401 | 19.601 | invalid | 6 |
| V | Z | AH1 | T | 1 | 15.806 | 19.601 | invalid | 6 |
| ZH | L | ER1 | K | 2 | 33.442 | 5.678 | invalid | 5 |
| SH | F | IY1 | D | 1 | 23.258 | 3.701 | invalid | 5 |
| Z | N | AE1 | T | 1 | 25.541 | 5.678 | invalid | 4 |
| F | N | EH1 | T | 2 | 23.969 | 3.315 | invalid | 3 |
| F | K | IY1 | P | 1 | 23.905 | 3.685 | invalid | 3 |
| V | T | IH1 | N | 2 | 22.639 | 3.685 | invalid | 3 |
| Z | V | IY1 | L | 2 | 26.018 | 15.023 | invalid | 3 |
| Z | M | AE1 | T | 1 | 21.983 | 5.678 | invalid | 2 |
| ZH | M | AE1 | T | 1 | 26.800 | 5.678 | invalid | 2 |
| F | M | AE1 | T | 4 | 21.800 | 3.315 | invalid | 2 |
| SH | P | EY1 | L | 2 | 26.172 | 0.000 | invalid | 2 |
| V | M | AE1 | T | 2 | 20.388 | 3.315 | invalid | 1 |
| V | N | EH1 | T | 2 | 24.017 | 3.315 | invalid | 1 |
| SH | K | IY1 | P | 2 | 26.976 | 0.000 | invalid | 1 |
| Z | P | EY1 | L | 1 | 25.421 | 5.678 | invalid | 1 |
| ZH | P | EY1 | L | 1 | 32.906 | 5.678 | invalid | 1 |
| ZH | T | IH1 | N | 1 | 29.763 | 5.678 | invalid | 1 |
| ZH | K | IY1 | P | 1 | 33.710 | 5.678 | invalid | 1 |
| ZH | N | EH1 | T | 1 | 33.775 | 5.678 | invalid | 0 |
| Z | K | IY1 | P | 1 | 27.547 | 5.678 | invalid | 0 |
| V | P | EY1 | L | 2 | 25.782 | 3.685 | invalid | 0 |
| V | K | IY1 | P | 1 | 26.586 | 3.685 | invalid | 0 |
| ZH | V | IY1 | L | 1 | 32.181 | 15.023 | invalid | 0 |

# B English syllabification

For every entry in the CELEX database, there is a corresponding broad syllabified transcription of the word in a Received Pronunciation accent. This appendix describes an automated procedure used to process these transcripts and to separate medial clusters from their flanking nuclei, parsing the resulting sequences into coda and onset, and reversing allophonic processes targeting medial clusters.

While the segmental content of these transcriptions is precise, the CELEX syllabifications are unsystematic. Given the absence of contrastive syllabification in English (if not all languages: see Blevins 1995:221, Elfner 2006), any sequence of a medial consonant cluster and its flanking nuclei should receive the same syllabification in all words in which it occurs. This is not always the case with the CELEX transcriptions, however. For instance, the sequence [ɪstɪɪ] receives a different parse in *chemistry* ['kɛ.mɪ.stɪɪ] and *ministry* ['mɪ.nɪs.tɪɪ].[4] Consequently, these syllabifications are not used here.

## B.1 Ambiguous segments

The syllabification procedure begins by separating sequences of vocalic and consonantal segments. In English, *r* and onglides pattern with consonants or with vowels depending on the context in which they occur. The heuristic adopted here is that ambiguous segments which impose restrictions on adjacent vowels are themselves vocalic, and those which impose restrictions on adjacent consonants are consonantal.

Initially, between two vowels, or finally, *r* is consonantal. Before another consonant, however, *r* has been lost in Received Pronunciation. Even in *r*-ful dialects, though, post-vocalic non-onset *r* patterns with vowels, not coda consonants. Before non-onset *r* many vowel contrasts are suspended (e.g., Fudge 1969:269f., Harris 1994:255): compare American English *fern*/*fir*/*fur* to *pet*/*pit*/*putt*. In this position, *r* is the only consonant which permits variable glottalization of a following /t/ in *r*-ful British dialects (Harris 1994:258), and the only consonant which does not trigger variable deletion of a following word-final /t, d/ in American dialects (Guy 1980:8). This is shown in (2–3) below.

(2) /t/-Glottalization in *r*-ful British dialects:

| | | | |
|---|---|---|---|
| a. | des[ɚt] | ∼ | des[ɚʔ] |
| | c[ɚt]ain | ∼ | c[ɚʔ]ain |
| b. | fi[st] | ∼ | *fi[sʔ] |
| | mi[st]er | ∼ | *mi[sʔ]er |

(3) /t, d/-Deletion in American English:

| | | | |
|---|---|---|---|
| a. | be[lt] | ∼ | be[l] |
| | me[nd] | ∼ | me[n] |
| b. | sk[ɚt] | ∼ | *sk[ɚ] |
| | th[ɚd] | ∼ | *th[ɚ] |

Following Pierrehumbert (1994), pre-consonantal *r* is assigned to the preceding nucleus.

The front onglide is assigned to onset position when initial or preceded by a single consonant, as in [j]*arn* or *ju*[n.j]*or*. When the glide is preceded by two or more consonants, it is assigned to the nucleus. There is considerable evidence in support of this assumption.

[4] Note that word-final *y* is usually lax in Received Pronunciation (Wells 1982:II.294).

When [j] is assigned to the onset, it may be followed by any vowel (Borowsky 1986:276), but when it is nuclear, the following vowel is always [u], suggesting a nuclear affiliation (Harris 1994:61f., Hayes 1980:232). Clements and Keyser (1983:42) note that [j] is the only consonant which can follow onset /m/ and /v/: [mj]*use*, [vj]*iew*. Finally, [ju] sequences in words such as *spew* behave as a unit in language games (Davis and Hammond 1995, Nevins and Vaux 2003) and speech errors (Shattuck-Hufnagel 1986:130).[5]

The phonotactic properties of the back onglide [w] are quite different than those of the front onglide, and it is consequently assigned to the onset portion of medial clusters. Whereas [j] shows only limited selectivity for preceding tautosyllabic consonants (Kaye 1996), [w] only rarely occurs after onset consonants other than [k] (e.g., *tran*[kw]*il*), and never after tautosyllabic labials in the native vocabulary. Whereas [kj] is always followed by [u], [kw] may precede nearly any vowel (Davis and Hammond 1995:161).

## B.2 Parsing medial consonant clusters

Medial consonant clusters are segmented into coda and onset using a heuristic version of the principle of onset maximization (e.g., Kahn 1976:42f., Kuryłowicz 1948, Pulgram 1970:75, Selkirk 1982:358f.) which favors parses of word-medial clusters in which as much of the cluster as possible is assigned to the onset. A medial onset is defined to be "possible" simply if it occurs word-initially (according to the rules defined above). As an example, the medial clusters in words such as *neu*[.tɹ]*on* or *bi*[.stɹ]*o* also occur in word-initial position (e.g., [tɹ]*ain*, [stɹ]*ike*), so the entire cluster is assigned ot the onset. In contrast, the cluster in *mi*[n.stɹ]*el* is not found word-initially; the maximal onset here is [stɹ] and the remaining [n] is assigned to the preceding coda.

In English, when a medial consonant cluster is preceded by a stressed lax vowel, as *wh*[ɪs.p]*er*, *v*[ɛs.t]*ige*, or *m*[ʌs.k]*et*, the first consonant of the cluster checks the lax vowel (Hammond 1997:3, Treiman and Zukowski 1990). As Harris (1994:55) notes, however, when the medial cluster is also a valid onset, as in *whi*[s.p]*er*, *ve*[s.ti]*ge*, and *mu*[s.k]*et*, onset maximization will incorrectly assign the entire cluster to the onset and leave the lax vowel unchecked. For this reason, onset maximization parses are modified to assign the first consonant of a complex medial consonant cluster to the coda before a stressed lax vowel (Pulgram 1970:48).

## B.3 Phonologization

Following Pierrehumbert (1994), the traditional analysis of affricates as single segment (e.g., *SPE*:321f., Jakobson et al. 1961:24) rather than sequences of a stop and fricative (e.g., Hualde 1988, Lombardi 1990) is adopted here. In many languages, affricates pattern with simple onsets; for instance, Classical Nahua bans true onset clusters but permits the affricate series [ts, tʃ, tɬ] (Launey 2011:9). Other languages, such as Polish, distinguish affricates and stop-fricative sequences (Brooks 1965), providing further evidence that "true" affricates are represented as single segments (or single timing units), and in contrast with stop-fricative clusters (Clements and Keyser 1983:34f.).

In English, [ŋ] has been analyzed as a pure allophone of /n/ before underlying /k, g/ (with later deletion of /g/ in some contexts; Borowsky 1986:65f., *SPE*:85, Halle and Mohanan

[5]The glide is also assumed to be present in underlying representation (e.g., Anderson 1988, Borowsky 1986:278) rather than inserted by rule (e.g., *SPE*:196, Halle and Mohanan 1985:89, McMahon 1990:217) since presence or absense of the glide is contrastive (e.g., *booty*/*beauty*, *coot*/*cute*).

1985:62), or as a phoneme in its own right (e.g., Jusczyk et al. 2002, Sapir 1925). Onset [ŋ] is totally absent in onset position, where it cannot be followed by a /k, g/ needed to derive the velar allophone, a fact predicted only by the former account, and English speakers have considerable difficulty producing initial [ŋ] (Rusaw and Cole 2009). Following Pierrehumbert (1994), the allophonic analysis is assumed here. When followed by /k, g/, [ŋ] is mapped to /n/. When not followed by a velar stop (i.e., finally), [ŋ] is analyzed as underlying /ng/.

# References


Albright, Adam. 2007. Natural classes are not enough: Biased generalization in novel onset clusters. Ms., MIT.

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

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

Albright, Adam, Giorgio Magri, and Jennifer Michaels. 2008. Modeling doubly marked lags with a split additive model. In *Proceedings of the 32nd annual Boston University Conference on Lanugage Development*, volume 1, 36–47. Somerville, MA: Cascadilla.

Anderson, John M. 1988. More on slips and syllable structure. *Phonology* 5:157–159.

Anttila, Arto. 1997. Deriving variation from grammar. In *Variation, Change and Phonoloy Theory*, ed. Frans Hinskens, Leo Wetzels, and Roeland van Hout, 35–68. Amsterdam: John Benjamins.

Armstrong, Sharon L., Lila R. Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition* 13:263–308.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44:586–591.

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

Berent, Iris, and Joseph Shimron. 2003. Co-occurrence restrictions on identical consonants in the Hebrew lexicon: Are they due to similarity? *Journal of Linguistics* 39:31–55.

Berent, Iris, Joseph Shimron, and Vered Vaknin. 2001. Phonological constraints on reading: Evidence from the Obligatory Contour Principle. *Journal of Memory and Language* 44:644–665.

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

Berkley, Deborah M. 1994a. The OCP and gradient data. *Studies in the Linguistic Sciences* 24:59–72.

Berkley, Deborah M. 1994b. Variability in Obligatory Contour Principle effects. In *Papers from the 30th meeting of the Chicago Linguistic Society*, 1–12. Chicago: Chicago Linguistic Society.

Blevins, Juliette. 1995. The syllable in phonological theory. In *The handbook of phonological theory*, ed. John Goldsmith, 206–244. Oxford: Blackwell.

Borowsky, Toni. 1986. Topics in the lexical phonology of English. Doctoral dissertation, University of Massachusetts, Amherst. Published by Garland, New York, 1991.

Borowsky, Toni. 1989. Structure preservation and the syllable coda in English. *Linguistic Inquiry* 7:145–166.

Brooks, Maria Zagorska. 1965. On Polish affricates. *Word* 20:207–210.

Brown, Roger, and Donald Hildum. 1956. Expectancy and the perception of syllables. *Language* 32:411–419.

Brysbaert, Marc, and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41:977–990.

Cairns, Charles E. 1972. Review of Scholes 1966. *Foundations of Language* 9:135–142.

Chomsky, Noam. 1955. The logical structure of linguistic theory. Ms., Harvard University and MIT. Revised version published by Plenum, New York, 1975.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, Noam. 1986. *Barriers*. Linguistic Inquiry monographs. Cambridge: MIT Press.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. Cambridge: MIT Press.

Chomsky, Noam, and George A. Miller. 1963. Introduction to the formal analysis of natural languages. In *Handbook of mathematical psychology*, ed. R. Duncan Luce, Robert R. Bush, and Eugene Galanter, II.269–321. New York: Wiley.

Clements, George N., and Samuel Jay Keyser. 1983. *CV phonology: A generative theory of the syllable*. Cambridge: MIT Press.

Cleveland, William S., and Susan J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83:596–610.

Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84:218–257.

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

Coltheart, Max, Eddy J. Davelaar, Jon T. Jonasson, and Derek Besner. 1977. Access to the internal lexicon. In *Attention and performance VI*, ed. Stanislav Dornic, 535–555. Hillsdale, NJ: Lawrence Erlbaum.

Davidson, Lisa. 2005. Addressing phonological questions with ultrasound. *Clinical Linguistics and Phonetics* 19:619–633.

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

Davidson, Lisa. 2006b. Phonotactics and articulatory coordination interact in phonology: Evidence from non-native production. *Cognitive Science* 30:837–862.

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

Davis, Stuart, and Michael Hammond. 1995. On the status of onglides in American English. *Phonology* 12:159–182.

Dell, François, and Mohamed Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics* 7:105–130.

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

Elfner, Emily. 2006. Contrastive syllabification in Blackfoot. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 141–149. Somerville, MA: Cascadilla.

Fikkert, Paula. 1994. On the acquisition of prosodic structure. Doctoral dissertation, Leiden University.

Fowler, Carol A. 1987. Consonant-vowel cohesiveness in speech communication as revealed by initial and final consonant exchanges. *Speech Communication* 6:231–244.

Fowler, Carol A., Rebecca Treiman, and Jennifer Gross. 1993. The structure of English syllables and polysyllables. *Journal of Memory and Language* 32:115–140.

Frauenfelder, Ulrich H., R. Harald Baayen, Frauke M. Hellwig, and Robert Schreuder. 1993. Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language* 32:781–804.

Frisch, Stefan A., Nathan R. Large, and 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.

Fudge, Erik C. 1969. Syllables. *Journal of Linguistics* 5:253–286.

Gallagher, Gillian. In press. Speaker awareness of non-local ejective phonotactics in Cochabamba Quechua. *Natural Language and Linguistic Theory* to appear.

Gibson, Edward, and James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes* 14:225–248.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the workshop on variation within Optimality Theory, Stockholm University*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120. Stockholm: Stockholm University.

Gorman, Kyle. In press. Structural and and accidental phonotactic gaps. Paper presented at WCCFL 30 (to appear in the proceedings).

Greenberg, Joseph H., and James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English, I and II. *Word* 20:157–177.

Gruber, Robert P., and Richard A. Block. 2005. Effects of caffeine on prospective duration judgements of various intervals depend on task difficulty. *Human Psychopharmacology* 20:275–285.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating language in time and space*, ed. William Labov, 1–35. New York: Academic Press.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18:54–72.

Halle, Morris, and K. P. Mohanan. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16:57–116.

Hammond, Michael. 1997. Vowel quantity and syllabification in English. *Language* 73:1–17.

Harris, John. 1994. *English sound structure.* Cambridge: Blackwell.

Havlicek, Larry L., and Nancy L. Peterson. 1976. Robustness of the Pearson correlation against violations of assumptions. *Perceptual and Motor Skills* 43:1319–1334.

Hay, Jennifer, Janet Pierrehumbert, and Mary E. Beckman. 2004. Speech perception, well-formedness and the statistics of the lexicon. In *Phonetic interpretation: Papers in Laboratory Phonology VI*, ed. John Local, Richard Ogden, and Rosalind A.M. Temple, 58–74. Cambridge: Cambridge University Press.

Hayes, Bruce. 1980. A metrical theory of stress rules. Doctoral dissertation, MIT.

Hayes, Bruce. 2000. Gradient well-formedness in Optimality Theory. In *Optimality Theory: Phonology, syntax, and acquisition*, ed. Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, 88–120. Oxford: Oxford University Press.

Hayes, Bruce, and James White. In press. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* to appear.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Hooper, Joan. 1973. Aspects of natural generative phonology. Doctoral dissertation, University of California, Los Angeles.

Hualde, Jose Ignacio. 1988. Affricates are not contour segments. In *Proceedings of the 7th West Coast Conference on Formal Linguistics*, 143–157. Stanford: Stanford Linguistics Association.

Huang, James. 1982. Logical relations in Chinese and the theory of grammar. Doctoral dissertation, MIT.

Itô, Junko. 1989. A prosodic theory of epenthesis. *Natural Language and Linguistic Theory* 7:217–259.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1961. *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge: MIT Press.

Jusczyk, Peter W., Paul Smolensky, and Theresa Allocco. 2002. How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition* 10:31–37.

Jäger, Gerhard. 2007. Maximum entropy models and Stochastic Optimality Theory. In *Architectures, rules, and preferences: Variations on themes by Joan W. Bresnan*, ed. Annie Zaenen, Jane Simpson, Tracy H. King, Jane Grimshaw, Joan Maling, and Chris Manning, 467–479. Stanford: CSLI.

Kabak, Barış, and William J. Idsardi. 2007. Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech* 50:23–52.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Doctoral dissertation, MIT. Published by Garland, New York, 1980.

Kaye, Jonathan. 1996. Do you believe in magic? The story of s+C sequences. In *A festschrift for Edmund Gussmann*, ed. Henryk Kardela and Bogdan Szymanek, 155–176. Lublin: Lublin University Press.

Kessler, Brett, and Rebecca Treiman. 1997. Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language* 37:295–311.

Kirby, James P., and Alan C.L. Yu. 2007. Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In *Proceedings of the International Congress of the Phonetic Sciences XVI*, 1389–1392.

Kuryłowicz, Jerzy. 1948. Contribution à la théorie de la syllabe. *Bulletin de la Société Polonaise de Linguistique* 8:80–114.

Launey, Michel. 2011. *An introduction to Classical Nahuatl*. New York: Cambridge University Press.

Levelt, Clara, Niels O. Schiller, and Willem J.M. Levelt. 2000. The acquisition of syllable types. *Language Acquisition* 8:237–264.

Lipinski, John, and Prahlad Gupta. 2005. Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory and Language* 52:171–192.

Lombardi, Linda. 1990. The nonlinear organization of the affricate. *Natural Language and Linguistic Theory* 8:375–425.

Luce, Paul A., and Nathan R. Large. 2001. Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes* 16:565–581.

Luka, Barbara J., and Lawrence W. Barsalou. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52:436–459.

Marr, David. 1982. *Vision*. New York: Freeman.

McMahon, April. 1990. Vowel shifts, free rides and strict cyclicity. *Lingua* 80:197–225.

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

Myers, Scott. 1987. Vowel shortening in English. *Natural Language and Linguistic Theory* 5:485–518.

Nevins, Andrew, and Bert Vaux. 2003. Metalinguistic, shmetalinguistic: The phonology of shm-reduplication. In *Papers from the 39th meeting of the Chicago Linguistic Society*, 702–721. Chicago: Chicago Linguistic Society.

Newmeyer, Frederick J. 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot'. *Theoretical Linguistics* 33:395–399.

Noether, Gottfried E. 1981. Why Kendall tau? *Teaching Statistics* 3:41–41.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In *Experimental phonology*, ed. John J. Ohala and Jeri J. Jaeger, 239–252. Orlando: Academic Press.

Pan, Ning, and William Snyder. 2003. Setting the parameters of syllable structure in early child Dutch. In *Proceedings of the 27th annual Boston University Conference on Language Development*, 615–625. Somerville, MA: Cascadilla.

Pan, Ning, and William Snyder. 2004. Acquisition of /s/-initial clusters: A parametric approach. In *Proceedings of the 28th annual Boston University Conference on Language Development*, 436–446. Somerville, MA: Cascadilla.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: A study of triconsonantal clusters in English. In *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, ed. Patricia A. Keating, 168–188. Cambridge: Cambridge University Press.

Pitt, Mark A., and James M. McQueen. 1998. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* 39:347–370.

Pulgram, Ernst. 1970. *Syllable, word, nexus, cursus*. The Hague: Mouton.

Pylkkänen, Liina, Andrew Stringfellow, and Alec Marantz. 2002. Neuromagnetic evidence for the timing of lexical activation: An MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain and Language* 81:666–678.

Pylyshyn, Zenon. 1984. *Computation and cognition: Towards a foundation for cognitive science*. Cambridge: MIT Press.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104:192–233.

Rose, Sharon, and Lisa King. 2007. Speech error elicitation and cooccurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* 50:451–504.

Rusaw, Erin, and Jennifer Cole. 2009. Learning constraints that oppose native phonotactics from brief experience. Paper presented at the Mid-Continental Workshop on Phonology.

Sapir, Edward. 1925. Sound patterns in language. *Language* 1:37–51.

Scholes, Robert J. 1966. *Phonotactic grammaticality*. Berlin: Mouton.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Schütze, Carson T. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2:206–221.

Selkirk, Elisabeth O. 1982. The syllable. In *The structure of phonological representations*, ed. Harry van der Hulst and Norval Smith, 337–385. Dordrecht: Foris.

Shademan, Shabnam. 2006. Is phonotactic knowledge grammatical knowledge? In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 371–379. Somerville, MA: Cascadilla.

Shademan, Shabnam. 2007. Grammar and analogy in phonotactic well-formedness. Doctoral dissertation, University of California, Los Angeles.

Shattuck-Hufnagel, Stefanie. 1986. The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology Yearbook* 3:117–149.

Smith, Nelson V. 1973. *The acquisition of phonology: A case study*. Cambridge: Cambridge University Press.

Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:118–129.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgements. *Language* 87:274–288.

Sprouse, Jon, and Diogo Almeida. Submitted. Power in acceptability judgement experiments and the reliability of data in syntax. Ms., University of California, Irvine, and New York University, Abu Dhabi.

Stevens, Stanley S. 1946. On the theory of scales of measurement. *Science* 103:677–680.

Suárez, Lidia, Seok Hui Tan, Melvin J. Yap, and Winston D. Goh. 2011. Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review* 18:605–611.

Treiman, Rebecca. 1983. The structure of spoken syllables: Evidence from novel word games. *Cognition* 15:49–74.

Treiman, Rebecca. 1986. The division between onsets and rimes in English syllables. *Journal of Memory and Language* 25:476–491.

Treiman, Rebecca, Carol A. Fowler, Jennifer Gross, Denise Berch, and Sarah Weatherston. 1995. Syllable structure or word structure? Evidence for onset and rime units with disyllabic and trisyllabic stimuli. *Journal of Memory and Language* 34:132–155.

Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in Laboratory Phonology V: Acquisition and the lexicon*, ed. Michael Broe and Janet Pierrehumbert, 269–282. Cambridge: Cambridge University Press.

Treiman, Rebecca, and Andrea Zukowski. 1990. Towards an understanding of English syllabification. *Journal of Memory and Language* 29:66–85.

Vaden, Kenneth, Harry R. Halpin, and Gregory S. Hickok. 2009. Irvine phonotactic online dictionary. URL `http://www.iphod.com`.

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

Vitevitch, Michael S., and Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood density in spoken word recognition. *Journal of Memory and Language* 40:374–408.

Vitevitch, Michael S., and 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, and David Kemmerer. 1997. Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech* 40:47–62.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press. 3 volumes.

Wolf, Matthew, and John J. McCarthy. 2009. Less than zero: Correspondence and the null output. In *Modeling ungrammaticality in Optimality Theory*, ed. Curt Rice and Sylvia Blaho, 17–66. London: Equinox.