80th Annual LSA Meeting—Albuquerque, NM
7 Jan 2006

Adam Albright
albright@mit.edu

Gradient phonotactic effects: lexical? grammatical? both? neither?

1 Introduction

(1) A well-known effect: gradient acceptability of novel phonological strings ("wug words")

- "How good would . . . be as a word of English?"

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

(2) Relation between gradient acceptability and lexical statistics has been demonstrated in many domains

- Acceptability of novel words

  Greenberg and Jenkins (1964), Ohala and Ohala (1986), Coleman and Pierrehumbert (1997), Vitevitch, Luce, Charles-Luce, and Kemmerer (1997), Frisch, Large, and Pisoni (2000), Bailey and Hahn (2001), and others
- Likelihood of morphophonological alternations (Eddington 1996; Bybee 2001; Pierrehumbert 2002; Ernestus and Baayen 2003, etc.)
- Morphological productivity: (Bybee 1995; Albright 2002a; Albright and Hayes 2003, etc.)

☞ *But what is the mechanism? How do these results relate to phonological theory, and grammars?*

(3) Three general views

1. Grammar is categorical, but performance is gradient
   - The grammar defines what is *possible* in the language (grammatical vs. ungrammatical)
   - Acceptability tasks ask speakers how *probable* a sequence is, invoking a variety of different (non-grammatical) calculations
   - Performance & task effects create gradient effects beyond what the grammar cares about
2. Grammar itself is probabilistic and gradient
   - Phonological grammars encode not only what combinations are *possible*, but also which are *probable*
   - Grammaticality is not "all or nothing", but is a continuous function reflecting the probability of the sequences involved (Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000; Albright and Hayes 2003)
   - Gradient acceptability judgments mirror gradient grammaticality intuitions
3. There is no grammar
   - Assessing novel words involves assessing the degree of support from existing words
   - Assessing the acceptability of novel word = trying to recognize them as real words, and gauging the degree of support from existing words

(4) How do we distinguish among these possibilities?

- Simply demonstrating that there are gradient effects does not settle the issue
- We need principles that help us decide whether a particular gradient effect is a by-product of the task or of language use, or whether it constitutes learned knowledge that has been extracted from the data

(5) Goal of this study

- Sketch several different computationally implemented models of phonotactic well-formedness, each embodying a different theory of where gradience comes from, and what factors we expect to influence acceptability
  - ➢ Factors known to affect on-line performance (neighborhood density, token frequency, similarity) (Luce 1986; Newman, Sawusch, and Luce 1997; Luce and Pisoni 1998)
  - ➢ Factors that play a role in grammatical generalizations (natural classes, type frequency)
- Compare the performance of the models, using two datasets of gradient phonotactic acceptability judgments
  - ➢ Bailey and Hahn (2001), and Albright and Hayes (2003)
- Claim: data currently supports a *stochastic grammatical* approach to gradient acceptability
  - ➢ Although each model has its own strengths, a model stated in "grammatical" terms (probabilistic statements about legal sequences of natural classes) captures the broadest range of data
  - ➢ Models that do not incorporate such statements make unsubstantiated predictions

(6) Outline

- Sketch two different classes of models of gradient acceptability, and their computational implementations
- Compare the performance of these models on experimental data
- Consider the extent to which the results provide evidence for grammatical status of gradient acceptability

2 Lexical vs. combinatorial models of gradience

(7) Bailey and Hahn (2001): distinguish between two fundamentally different types of knowledge that could be used to decide about novel words

- Lexical knowledge
  - ➢ Speakers know the words of their language
  - ➢ Hearing a novel word activates a set of real words, while attempting lexical access
  - ➢ The more words it activates, and the more similar it sounds to them, the more plausible it is as a possible word
- Phonotactic knowledge
  - ➢ Speakers attend to combinatorial possibilities of different sounds in their language
  - ➢ Novel words are parsed into constituent sounds, and the likelihood of combinations is assessed
  - ➢ The more probable/"less illegal" the combinations are, the better the word sounds

(8) Why this distinction is useful

- Two simple endpoints in a spectrum of possible models, which make maximally distinct predictions about the factors that should influence gradient acceptability
- "Lexicon-only" model:
  - ➢ If gradient acceptability depends on consulting the lexicon, then factors that are known to play a role in lexical access should matter
  - ➢ Lexical (token) frequency, neighborhood density, etc.

- "Sequences only" model:
  - ➢ If gradient acceptability results from statements about possible sound combinations, then factors that are relevant in grammatical descriptions should place a role
  - ➢ Type frequency, natural classes, markedness (?), etc.

(9) Strategy:

- Construct models that use either lexical or sequential knowledge to predict the acceptability of novel sequences
- Test to what extent their ability to use different types of information helps their performance in modeling experimentally obtained ratings

2.1 Lexical models of gradient acceptability

(10) Simplest estimate of lexical support: neighborhood density

- Number of words that differ by *n* changes (substitutions, additions, deletions)
  - ➢ E.g., novel *droff* [draf] has (in some dialects) the neighbors *trough*, *prof*, *drop*, *doff*, *dross*
- Neighborhood density has been shown to play a role in a wide variety of effects—including lexical decision times, mishearings, phoneme identification, and, most relevantly, the acceptability of wug words
  - ➢ Greenberg and Jenkins (1964), Ohala and Ohala (1986), and others
- Generally too crude to model acceptability of novel words accurately, though, since even relatively "ordinary" wug words often have few or no neighbors
  - ➢ E.g., *drusp* [drʌsp], *stolf* [stɔlf], *zinth* [zɪnθ] all have 0 neighbors

(11) Bailey & Hahn (2001): point out that although such words have no immediate neighbors, they are fairly similar to many slightly more distant words

- E.g., *drusp* [drʌsp]: *trust*, *dusk*, *rusk*, *truss*, *dust*, etc.
- To capture this, we need to count words that are farther than a single substitution away, while at the same time paying attention to the severity of different substitutions

(12) Bailey and Hahn's Generalized Neighborhood Model (GNM)

Lexical support for a novel word = summed similarity of novel word to each existing word

- Exemplar model based on Nosofsky's Generalized Context Model (GCM), a similarity-based classification model (Nosofsky 1986; Nosofsky 1990)
- Every word in the lexicon contributes some amount of support, but very similar words contribute more
  - ➢ Similarity of words is assessed by finding minimum string edit distance (Kruskal 1983/1999; see also Bailey and Hahn 2001, Albright and Hayes 2003), using metric of segmental similarity based on shared natural classes (Frisch, Pierrehumbert, and Broe 2004) [1]
- Novel words are predicted to be better, the greater the number of words is that they are similar to, and the greater the similarity is to those words
  - ➢ Support for *zin* [zɪn] includes a number of similar words:

zen	0.609
sin	0.613
din	0.649
in	0.700
zing	0.720

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

➢ Support for *snulp* [snʌlp] is more distant, and similarity drops off more quickly:

snub	1.260
sulk	1.271
slump	1.283
snuff	1.367
null	1.400

- Words with higher lexical frequency also contribute more (according to a parameter)

2.2 Phonotactic models of gradient acceptability

(13) Phonological markedness constraints typically ban particular sounds, or particular combinations of sounds (cooccurrence restrictions)

- $*\begin{bmatrix} -\text{back} \\ +\text{round} \end{bmatrix}$ (no front rounded vowels)
- $*ti$
- $*\begin{bmatrix} +\text{nas} \\ \alpha\text{place} \end{bmatrix}\begin{bmatrix} -\text{son} \\ -\alpha\text{place} \end{bmatrix}$ (nasal place assimilation)

(14) A simple model of the probability of cooccurrence: *n*-gram models

- Bigram probability:
  - ➢ Probability of sequence $ab = \dfrac{\text{Number of times } ab \text{ occurs in the corpus}}{\text{Total number of sequences in the corpus}}$
- Bigram transitional probability:
  - ➢ Probability of b coming after $a = \dfrac{\text{Number of times } ab \text{ occurs in the corpus}}{\text{Total number of times } a \text{ appears before anything}}$
  - ➢ Probability of a string $abcd$ = P(initial a) × P(b after a) × P(c after b) × P(d after c)

(15) Numerous studies have shown effects of sequence probability on morpheme and word segmentation, age of acquisition, recall of nonwords, and, most relevant to the current study, gradient acceptability of wug words

- Coleman and Pierrehumbert (1997): log likelihood of most probable parse into onset/rhyme constituents
- Frisch, Large, and Pisoni (2000): onset-rhyme frequencies
- Vitevitch and Luce (2004), Vitevitch and Luce (2005): positional bigram frequency
  - ➢ Bigrams counted separately for different positions in the word, and words are weighted by token frequency

(16) In principle, we could count trigrams, tetragrams, etc. However, for nonce words, we encounter an issue akin to the "no neighbors" problem in (11) above

- *dresp* [drɛsp]: there are no words containing [ɛsp] in English, yet native speakers tend to rate this word as relatively acceptable (4.7 on a scale of 1 to 7; mean=4.0, max=6.5)
- More generally: for larger values of *n*, the number of possible *n*-grams grows exponentially, meaning a much larger corpus is needed to estimate their frequencies accurately
- In this case, our corpus is the lexicon. Since lexicons are of a limited size, we inevitably end up with many accidental gaps.
  - ➢ I.e., we can't get more data about whether [ɛsp] is legal by simply collecting more words
- Many different strategies have been developed to handle this problem (see, e.g., Jurafsky and Martin 2000), mostly by combining information from shorter and longer values of *n*
- A strategy of interest to phonologists, though, is to try to reason about well-formedness of underattested sequences based on *natural classes*

(17) The intuition

- Although the [ɛsp] in *dresp* is unattested, it does get support from other similar sequences
  - ➢ [ɪsp] (*crisp, wisp, lisp, . . .*), [æsp] (*clasp, rasp, asp, . . .*))
  - ➢ [ɛst] (*best, west, rest, . . .*), [ɛsk] (*desk*)
- Taken together, they suggest that sequences of [*lax vowel* + s + *voiceless stop*] are allowed in English (Hammond 1999, p. 115)
- To discover this, we must consider not just of particular segments, but also classes of segments

(18) Comparing sequences to extract more general patterns of natural classes

- The Minimal Generalization approach (Albright and Hayes 2002; Albright and Hayes 2003)

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

(19) Not all comparisons yield equally illuminating generalizations, however

	æ	s	p
+	b	o	a
→	[+voi]	*seg*	*seg*

- Assuming no feature values shared between [s]∼[o] or [p]∼[a], we simply learn that voiced segments can be followed by two more segments
- In fact, combinations of [+voi] + two more segments are *extremely* well attested in English!
- Potentially fatal prediction: *bzarshk* [bzarʃk] should be very acceptable, because [bza] and [rʃk] get lots of support from other [+voi] *seg seg* sequences

(20) The challenge

- Find a way to count over natural classes such that [ɪsp] and [æsp] provide strong support for [ɛsp], while [æsp] and [boa] provide little or no support for [bza] or [rʃk]

(21) The leading idea

- Intuitively, [ɪsp] + [æsp] → [ɛsp] is much less of a leap; they practically make [ɛsp] inevitable
- This is due to the fact that [−back, −round, −tense] *sp* is so specific
  - ➢ All we need to do is specify the vowel height, and we have [ɛsp]
  - ➢ To get [ɛsp] from [+voi] *seg seg*, we need to fill in very many features
- Put differently: [−back, −round, −tense] *sp* describes a small set of possible sequences (ɪsp, ɛsp, æsp)
  - ➢ If such sequences are legal, the probability of finding any one of them at random is 1/3
  - ➢ The set of [+voi] *seg seg* sequences is huge; the chance of getting [ɪsp] or [æsp] at random is tiny
  - ➢ Although both [−back, −round, −tense] *sp* and [+voi] *seg seg* can describe sequences like [ɪsp] or [æsp], the former characterization makes it much more likely that we would encounter these particular sequences

(22) *Maximum likelihood estimation* (MLE)

- Find the description that makes the training data as likely as possible
- "English words conform to certain shapes because they have to, not out of sheer coincidence"
- Related to OT ranking principles that seek the most restrictive possible grammar (Prince and Tesar 2004; Hayes 2004); also related to Bayesian inference
- For an MLE-based approach to *n*-gram models that refer to classes of items, see Saul and Pereira (1997)[2]

(23) Implementation: instantiation costs

- Simple bigram model:
  - ➢ Probability of sequence $ab = \frac{\text{Number of times } ab \text{ occurs in corpus}}{\text{Total number of two-item sequences}}$
- Stated over natural classes:
  - ➢ Probability of sequence *ab*, where $a \in$ class *x*, $b \in$ class *y*

    $= \frac{\text{Number of times } xy \text{ occurs in corpus}}{\text{Total number of two-item sequences}} \times \text{Prob(choosing } a \text{ from } x) \times \text{Prob(choosing } b \text{ from } y)$
- What is the probability of any particular instantiation *a* of a natural class *x*?
  - ➢ Simple: $\frac{1}{\text{Size of } x \text{ (i.e., number of members)}}$
  - ➢ Weighted: Relative frequency of $a \times \frac{1}{\text{size of } x}$
- ☞ For reasons I don't have space to discuss here, it appears that weighting by (type) frequency is useful; this is what I will assume in the simulations reported here

(24) Example: probability of [ɛsp]

- Probability of [ɛsp] using trigram $\begin{bmatrix} -\text{back} \\ -\text{round} \\ -\text{tense} \end{bmatrix} sp$

  $= \text{Prob}(\begin{bmatrix} -\text{back} \\ -\text{round} \\ -\text{tense} \end{bmatrix} sp) \times \text{Prob([ɛ] among lax front vowels})$ (*relatively high*)
- Probability of [ɛsp] using trigram [+voi] *seg seg*

  $= \text{Prob([+voi] } seg\ seg) \times \text{Prob([ɛ] among voiced)} \times \text{Prob([s])} \times \text{Prob([p])}$ (*very low*)
- Given multiple possible ways to parse the same string of segments, find the one with the highest probability (Coleman and Pierrehumbert 1997; Albright and Hayes 2002)
  - ➢ [ɛsp] can find good support from $\begin{bmatrix} -\text{back} \\ -\text{round} \\ -\text{tense} \end{bmatrix} sp$
  - ➢ [bza] has no allies that provide such a close fit; it must rely on broader (and weaker) generalizations like [+voi] *seg seg*

(25) Local summary

- A method of evaluating sequences of natural classes, to determine which are supported by the training data
- Result: a set of (stochastic) statements about relative likelihood of different sequences

[2] Unfortunately, n-gram models based on classes of elements in syntax tend to assume that each item belongs to ideally one, or exceptionally a few, classes (*tree* is a verb, *of* is a preposition, etc.). For phonological applications, we need to consider each segment as a member of many classes simultaneously, and even a single instance of a segment may be best characterized in different terms with respect to what occurs before it vs. after it. Ultimately, we seek a ranking algorithm that incorporates the principle of MLE, without making the simplifying assumptions of existing class-based *n*-gram models.

3 Testing the models

(26) Full set of models considered here

- Lexical models
  - ➢ Traditional neighborhood density model ("number of neighbors"), with or without frequency weighting
  - ➢ Generalized Neighborhood Model (GNM), with or without frequency weighting
- Sequential models
  - ➢ Standard bigram, trigram transitional probability
  - ➢ Average bigram, trigram frequency
  - ➢ Sequential model based on bigrams on natural classes, with instantiation costs, described in (17)-(24) above
- "Hybrid" model (for comparison)
  - ➢ Vitevitch and Luce (2004): positional bigram frequencies, weighted for lexical frequency
  - ➢ http://www.people.ku.edu/~mvitevit/PhonoProbHome.html

☞ In part, a replication of Bailey & Hahn (2001), to allow direct comparison with results from other studies, and to compare models that they did not consider

(27) Training the models: input file containing all word forms with non-zero frequency in CELEX (Baayen, Piepenbrock, and van Rijn 1993), in phonemic transcription

- Also tried just lemmas, or just monosyllables; both tend to yield slightly worse results for tasks described below
- Since CELEX has many "duplicate" entries (same word broken up into two separate entries), and their token frequencies were combined
- CELEX uses British pronunciation; could impact ability to model results from American speakers for some sequences, but doesn't appear to be an issue here
- Vitevitch & Luce model uses its own (smaller) training set and frequencies

Models were then used to derive predictions for novel (wug) words, from two ratings experiments described in the literature

3.1 Dataset 1: Bailey and Hahn (2001)

(28) Bailey and Hahn (2001)

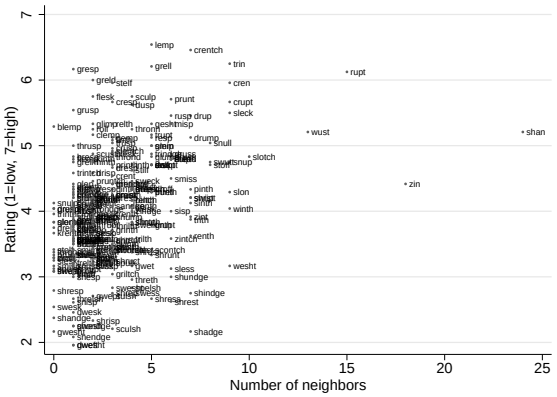
- Set of 259 monosyllabic non-words, of generally moderate acceptability
  - ➢ E.g., *drolf* [drɒlf], *smisp* [smɪsp], *pruntch* [prʌnt͡ʃ], *stulf* [stʌlf], *zinth* [zɪnθ], *glemp* [glɛmp]
  - ➢ Designed not to contain any overt phonotactic violations, but a handful of words did contain questionable sequences ([ʃt#] in *gwesht, swesht*, etc.; sNVN in *smimp*; etc.)
- Words presented auditorily in a carrier sentence ("*Zinth.* How typical sounding is *zinth*?")[3]
- Participants rated words on a scale from 1 (low) to 7 (high)
  - ➢ Ratings were then normalized statistically (see Bailey & Hahn 2001 for details)

[3] A written version of the experiment was also carried out, but I consider here only the oral task.

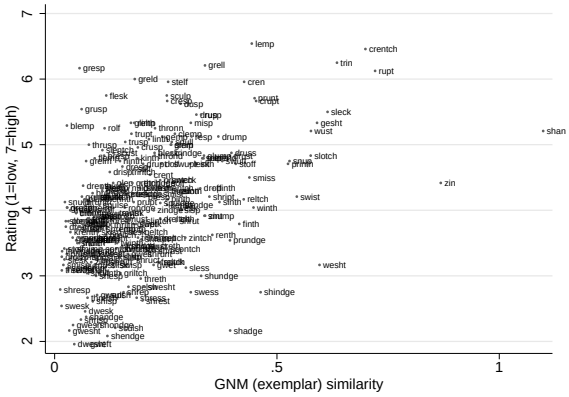
(29) Results at a glance

a. Lexical models

- Number of neighbors ($r = 0.385$)

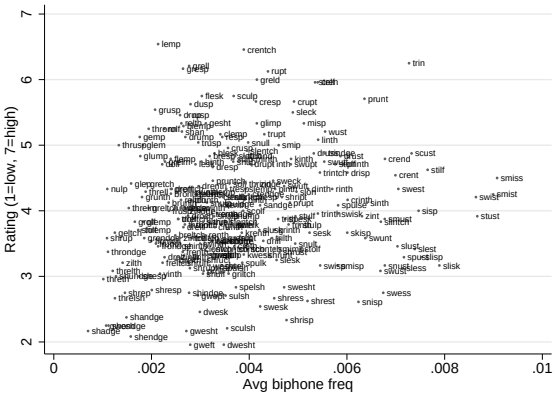


- GNM ($r = 0.455$)

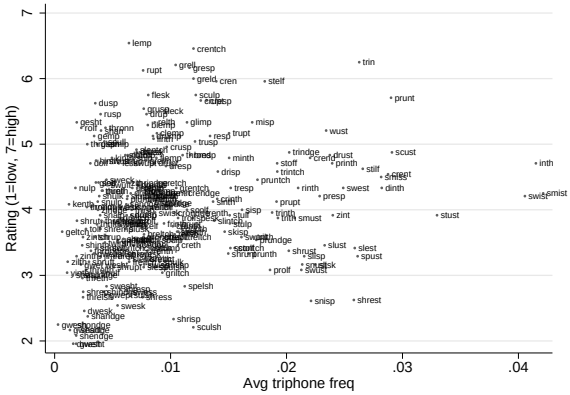


b. Sequential models

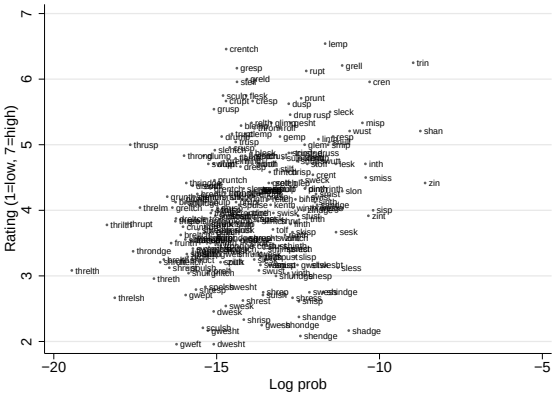
- Average biphone freq ($r = 0.125$)



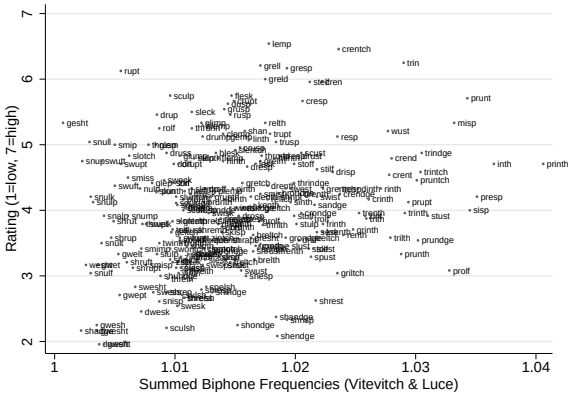
- Average triphone freq ($r = 0.195$)



- Natural class-based likelihood ($r = 0.296$)



- Vitevitch & Luce (2004) ($r = 0.234$)



- Transitional probabilities score substantially worse (not shown)

(30) Highlights: somewhat inconclusive

- All models capture a certain amount of the variance, though no model does extremely well
- GNM is overall best (though some of its numerical success comes from sparse group of outliers)

3.2 Dataset 2: Albright and Hayes (2003)

(31) Albright and Hayes (2003)

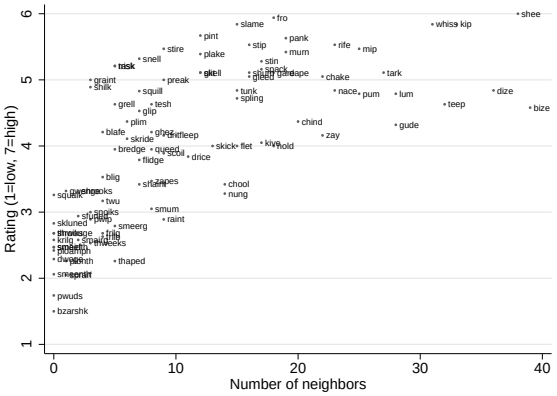
- 92 wug words used in pre-test, as control for a past tense study
- 62 items estimated ahead of time to be relatively acceptable; 30 "foils", at varyingly lesser degrees of acceptability (full set in Appendix)
  - ➢ Moderate to high acceptability: *kip* [kɪp], *stire* [staɪr], *pank* [pæŋk], *fleep* [flip], *blafe* [bleɪf]
  - ➢ Some marginal sequences:
    [pwʌdz] (*pw)
    [θɹɔɪks] (*[ɔɪk])
    [ʃwuʒ] (*ʃw)
    [skɪk], [snʌm] (*sC$_1$VC$_1$, *sNVN)
  - ➢ A few items with unattested rhymes: [smɛrg], [smiːlθ]
  - ➢ One very ill-formed word: [bzarʃk], used during training as an example of a word that most English speakers feel would not be a possible word
- Words presented auditorily in random order, in a carrier sentence ("*Blafe*. I like to *blafe*[4].")
- Participants repeated word aloud, and rated on a scale from 1 ("impossible as an English word") to 7 ("would make a fine English word")
  - ➢ If word was repeated incorrectly, rating for that trial was excluded from analysis
- 20 participants (1 excluded for too many incorrect repetitions)
  - ➢ High level of agreement! Correlation between participants 1-10 vs. 11-19: $r = 0.864$

[4] All words in this study were presented as verbs. It is not clear what effect this had on the final results.

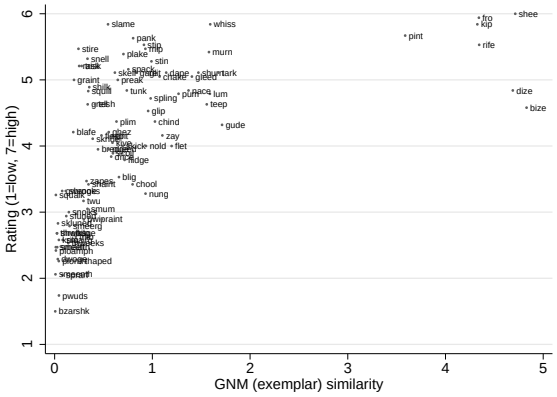
(32) Results at a glance

a. Lexical models

- Number of neighbors ($r = 0.675$)

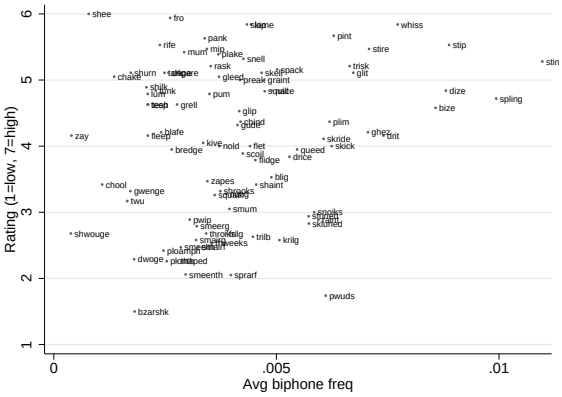


- GNM ($r = 0.560$)

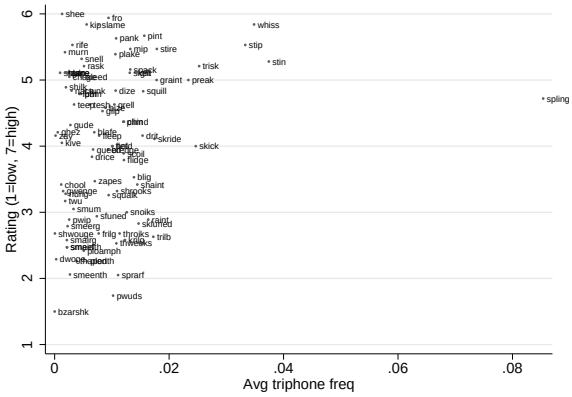


b. Sequential models

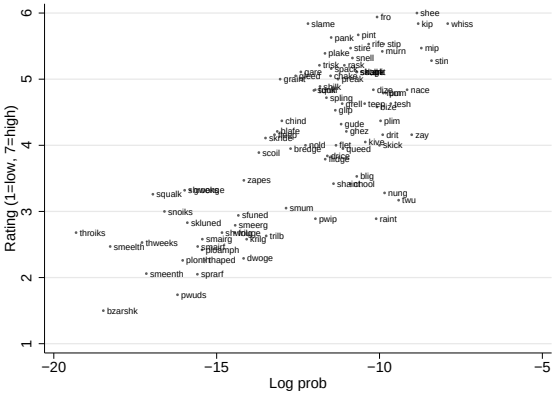
- Average biphone freq ($r = 0.191$)



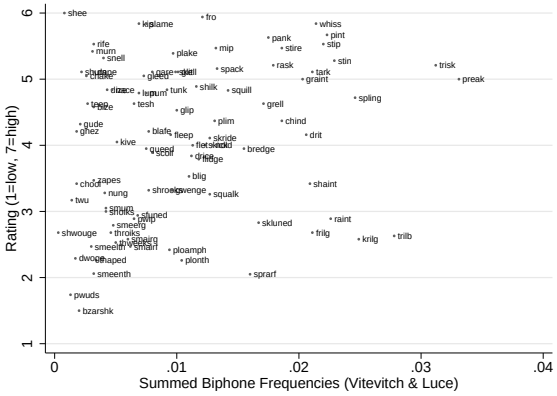
- Average triphone freq ($r = 0.215$)



- Natural class-based likelihood ($r = 0.757$)



- Vitevitch & Luce (2004) ($r = 0.216$)



(33) Points to note

- In this task, the model based on natural classes comes out significantly ahead
  - ➢ Not only the best model for this task, but overall the cleanest result of the whole study
- Simple NNB model also surprisingly effective! However, only when token frequency is not taken into account (more below)

3.3 Discussion

(34) Summary of modeling results over both datasets

- Bailey & Hahn: lexical models show better performance, but no model does particularly well
- Albright & Hayes: sequential model based on natural classes comes out ahead (and does quite well)

(35) Questions to be addressed

- Why do all the models do so poorly on the Bailey & Hahn data?
- Why does a model based on possible combinations of natural classes do better on the Albright & Hayes data?
- What does this tell us about the nature of gradient acceptability?
- What kinds of data are still needed to settle the issue?

(36) Why does the Bailey & Hahn data set appear to be so noisy?

Two possible reasons for the difference:

- Differences in tasks
  - ➢ Participants did not repeat words (can't flag mishearings, less incentive to pay attention)
  - ➢ Instructions asked a different question: "How typical sounding is . . . " (as opposed to "how possible is . . . as an English word")
- Differences in wug words
  - ➢ Bailey & Hahn items were all of intermediate acceptability (low: *gweft*; high: *crendge*)
  - ➢ Albright & Hayes items involved training on "endpoints" (low: *bzarshk*; high: *kip*), and included more variability between test items themselves
  - ➢ Perhaps this variability helps anchor endpoints, so participants can use the scale in a more meaningful/consistent way?

☞ Possible lesson for future studies

- ➢ Given that unlikely words are frequently "repaired" in perception, having participants repeat words seems to be an important check
- ➢ Bailey and Hahn use intermediate items for statistical reasons (see their discussion on this point), but a compromise is possible
  - ◦ Large number of intermediate items, but including enough "endpoint" items to anchor the scale

(37) A more important reason why it is important to include items with phonotactic violations

- Illegal sequences (cooccurrence restrictions, or constraints on possible combinations of sounds) are at the heart of what phonology traditionally aims to explain
- Even if an exemplar model does very well at predicting that *kip* sounds better than *shresp*, if it can't explain why *dlap* or *mrut* are worse than any of these intermediate words, we can't claim to have gotten very far
- This is significant, because similarity-based models are, on the face of it, rather ill-suited for capturing facts about violations

(38) An example: [sruː]

- The nonwords [fruː] and [sruː] both have many neighbors (*brew*, *crew*, *drew*, *grew*, *roux*, *screw*, *shrew*, etc.), but [sruː] contains a phonotactic violation (*sr)
- GNM predictions for these words:

fruː	1.96
sruː	1.68

  - ➢ Point of comparison: [lʌm] 1.59, [wɪs] 1.59, [tark] 1.68; all rated >5 out of 7 in Albright & Hayes study

- I suspect that this highly overestimates the goodness of [sruː]

☞ Forms like this are probably just the tip of the iceberg! Data like that in Bailey & Hahn (2001) allow us to probe only a tiny part of the overall picture of how well models account for gradient acceptability

(39) So why does a sequential model based on natural classes do better on the Albright and Hayes data?

- This dataset contains some words with (mild) phonotactic violations
- The 20 forms from the Albright & Hayes dataset which the GNM most seriously *overestimates*:
  - ➢ θɪɔɪks, ʃwuʒ, raɪnt, frɪlg, krɪlg, smɛɪg, trɪlb, smiːlθ, smɛɪf, θwiks, ploʊmf, dwoʊd͡ʒ, ploʊnθ, daɪz, θeɪpt, sminθ, spɹaɪf, baɪz, pwʌdz, bzaɪʃk
  - ➢ The majority of these contain some type of violation, which the GNM is not able to pick up on
- The 20 forms that the GNM most seriously *underestimates*
  - ➢ sleɪm, staɪɪ, pæŋk, snɛl, ræsk, trɪsk, stɪp, pleɪk, mɪp, wɪs, graɪnt, skɛl, spæk, stɪn, ʃɪlk, skwɪl, gɛɪ, priːk, glɪt, mr̩n
  - ➢ These are mostly fine sequences, which just happen to be a bit isolated in the lexicon

(40) Summarizing the discussion thus far:

- Poor performance on the Bailey & Hahn items seems to stem from the fact that it's overall a rather messy dataset, perhaps because of the way the experiment was set up
- In a cleaner data set with a wider range of items, we see that the ability to pay attention to possible sequences is important in modeling human intuitions

☞ Gradient acceptability reflects knowledge about the relative probability of different combinations of natural classes, not knowledge of words directly

(41) One last "unlexical" effect

- None of the models explored here derived any advantage from their ability to take token frequency into account
  - ➢ GNM and simple NNB models did best when frequency weighting was turned off
  - ➢ Vitevitch and Luce model, which has frequency weighting built in, does not come out ahead because of it
- An apparent difference from Bailey and Hahn (2001), who found significant contribution of token frequency
  - ➢ However, even there, only a tiny numerical boost was observed (r^2 gain of .01?)
  - ➢ I was unable to replicate this effect, even on their data
- In most cases, taking token frequency into account doesn't change the predictions at all
  - ➢ Most words in the lexicon are very low frequency, so a boost for high token frequency only gives more influence to a small set of words
- When it does make a difference, though, it tends to be a deleterious one
  - ➢ Compared predictions of GNM with and without token frequency, to find those words which changed the most when token frequency was considered

    [zeɪ] (*say*), [gɛr] (*there*), [paɪnt] (*mine*), lʌm (*come*), [gliːd] (*need*)
  - ➢ For all five of these words, the frequency-sensitive model was *farther from* the human ratings (overestimated goodness)
  - ➢ Further evidence that pattern strength is related to type, not token frequency (Bybee 1995; Albright 2002b; Albright and Hayes 2003; Hay, Pierrehumbert, and Beckman 2004)

(42) I take this finding to be at odds with the idea that gradient acceptability arises as a by-product of consulting the lexicon.

☞ Lexical access is known to be highly sensitive to frequency, yet gradient acceptability appears to be completely impervious to it

4 Conclusion

(43) Based on currently available data, it appears that gradient phonotactic acceptability bears the markings of a grammatical effect

- It seems to reflect knowledge about sequences, stated in terms of natural classes
- It is not sensitive to lexical frequency, unlike other known performance effects

(44) Many respects in which the current data is inadequate to truly compare these different models

- Data sets do not contain sufficient information to calibrate the scale of unacceptability (including "truly unacceptable" sequences
- They don't contain items specifically designed to test the possible contribution of token frequency
  - ➢ Post-hoc comparisons appear to show no effect, but this is based on relatively few items
- Data from monosyllables is unable to test a whole range of questions about higher level prosodic restrictions

(45) Two immediate goals

- Experimental studies expanding the range of wug words that the model can be tested on, and comparing different sources of evidence (ratings, reaction times, incorrect repetitions, etc.)
- Modeling sequential restrictions by constraint ranking, in OT-theoretic grammar

5 Appendix: list of wug words from Albright & Hayes data set

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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