Adam Albright
albright@mit.edu

LSA 82—Chicago
3 Jan 2007

From clusters to words: grammatical models of nonce-word acceptability

1 Introduction

(1) Experimental investigation of phonotactic acceptability using nonce words

- The “blick” test (Chomsky and Halle 1965): ✓[blɪk] vs. *[bnɪk]
- Numerical acceptability ratings: how plausible is [blɪk] as a word of your language?

  7 = could easily be a word (very acceptable)
  4 = not perfect, but conceivable (intermediate)
  1 = completely implausible (very unacceptable)

- A consistent finding
  - ○ When nonce words are varied to contain more or less probable combinations of sounds, acceptability ratings systematically reflect these differences

    (Greenberg & Jenkins 1964; Ohala & Ohala 1986; Coleman & Pierrehumbert 1997; Treiman et al. 2000; Frisch, Large, & Pisoni 2000; Frisch & Zawaydeh 2001; Berent, Everett, & Shimron 2001; Bailey & Hahn 2001; Hammond 2004; Shademan 2007; etc.)

(2) Usual strategy: factor-based analysis

- Substantial success demonstrating gradient preferences by focusing on particular pre-selected aspects of nonce words
  - ○ Co-occurrence of adjacent consonants in clusters (medial: Hay, Pierrehumbert, and Beckman 2004; initial: Hayes and Wilson, in press)
  - ○ Co-occurrence probability of nucleus-coda combinations (Treiman, Kessler, Knewasser, Tincoff, and Bowman 2000)
  - ○ Non-local consonant-to-consonant co-occurrence probabilities (Frisch and Zawaydeh 2001; Berent, Everett, and Shimron 2001; Koo and Oh, this session)
  - ○ Etc.: onset-vowel and onset-tone combinations (Kirby and Yu 2007)
- Vary structure of interest, keep rest of word relatively constant
  - ○ Valuable in demonstrating reality of individual constraints
  - ○ Ignores other sources of variability that are irrelevant for comparison of interest

(3) The challenge: from subparts to whole words

- How are assessments of different aspects of words combined into a single rating?
- That is, how do individual constraints interact in the evaluation of entire words?
- A version of this question has been tackled by several previous studies (Ohala and Ohala 1986; Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000), but generally not with items selected specifically to systematically distinguish predictions of different models

(4) Goals and outline

- Systematic empirical examination of how probabilities in different parts of the word interact, focusing on a more confined domain
  - ○ Simultaneous contribution of onsets and rhymes to ratings of monosyllables
- Compare two general approaches
  - ○ Cumulative models (harmonic grammar, stochastic parsing)
  - ○ Non-cumulative "fatal violation" models (OT)
  - ○ Claim: cumulative models provide simplest and most accurate account
  - ○ Echoes similar claims by Coleman and Pierrehumbert (1997), Frisch, Large, and Pisoni (2000), Pater (2007)
- Outline
  - ○ Background and outline of the experimental data
  - ○ Modeling simultaneous contribution of violations in multiple parts of the word
  - ○ Modeling ganging up effects

2 Cumulative vs. non-cumulative evaluation

(5) Standard interpretation of OT: candidates are as bad as their worst violation

- Example: evaluation of [bnɪk], [bnælθ] and [sfælθ] as hypothetical words of English
- Schematic illustration
  - ○ Initial *#bn* is ungrammatical, typically repaired by epenthesis: *[bn ≫ DEP
  - ○ Initial *#sf* and final *lθ* are grammatical (*sphere*; *health, filth*): DEP ≫ *lθ], *[sf

			*[bn	DEP	*lθ]	*[sf
	a.	bnɪk	*!			
☞	a.′	bənɪk		*		
	b.	bnælθ	*!		*	
☞	b.′	bənælθ		*	*	
☞	c.	sfælθ			*	*
	c.′	sfæləθ		*!	*	

(6) Predictions of OT, as standardly construed

- Nonce words [bnɪk], [bnælθ] are predicted to be equally ungrammatical, since both are eliminated by the same fatal violation of *[bn (candidates (6a), (6b))
  - ○ We consider below alternative models of candidate comparison (Berent, Everett, and Shimron 2001; Coetzee 2004)
- Ungrammatical nonce word [bnɪk] should be categorically worse than [sfælθ], which contains two rare but occurring clusters (candidates (6a) vs. (6c))

(7) Compare with linear model using summed weights, rather than strict ranking (Goldwater & Johnson 2003; Jäger, in press; Hayes & Wilson, in press; Pater, Bhatt, & Potts 2007)

- Constraints are given numerical weights, rather than relative ranks

- Each candidate is assigned a harmony score based on weighted sum of its violations
- Candidate with smallest penalty (highest harmony) wins

			*[bn	Dep	*lθ]	*[sf	Sum
		Weight:	−2.5	−2	−1.5	−1.5	
	a.	bnɪk	1				−2.5
☞	a.′	bənɪk		1			−2
	b.	bnælθ	1		1		−4
☞	b.′	bənælθ		1	1		−3.5
☞	c.	sfælθ			1	1	−3
	c.′	sfæləθ		1	1		−3.5

- Since violations are summed across the entire word, additional (non-fatal) violations of lower-weighted constraints contribute to predicted well-formedness[1]
  - [bnɪk] (penalty = −3) ≻ [bnælθ] (penalty = −4)
- Multiple violations of lower-ranked constraints can “gang up” and overcome single violations of higher-ranked constraints
  - [bnɪk] (penalty = -2.5) ≻ [sfælθ] (penalty = -3)

(8) Preliminary evidence in support of cumulative evaluation

- “Anti-bottleneck effects”: added influence of violations beyond the fatal violation
  - Ohala and Ohala (1986): asked subjects to compare pairs of nonce words sharing same rare/illegal structure, but differing in presence of a second violation
  - Subjects were asked to judged which was “closer to English”
  - Relatively strong preference for nonce words with fewer *overall* violations

One violation	Two violations	Preference for 1 violation
mlit	**ml**ɔ**ʒ**	73%
spøf	**ml**øf	94%
xrit	**x**lo**x**	69%
sfɛt	**sfʊb**	94%

- “Ganging up effects”: multiple violations of violable constraints may outweigh single violations of an inviolable constraint
  - Coleman and Pierrehumbert (1997): subjects judged [mɹuˈpeɪʃən] more acceptable than [ˈsplɛtɪsak], despite illegal initial *#mr*
  - Suggests that several somewhat low-probability constituents make a word less acceptable than one very low-probability constituent

(9) These data are suggestive, but far from conclusive

- Existing demonstrations of cumulativity are based on just a few items, or items that differ in many respects simultaneously
  - Although (or because?) cumulativity effects are so intuitive, they have not been demonstrated systematically with items designed specifically to test for them

[1] I provisionally assume, following Keller (2000, 2006) that the acceptability of a output candidate in a ratings task can be related straightforwardly to its harmony score. Boersma (2004) points out that this assumption may be problematic in modeling certain types of comparisons, and Pater (2007) proposes a comparative evaluation based on the difference between the harmony of the candidate under consideration and the best competing candidate for the same input. Unfortunately, this proposal does not suffice to distinguish the nonce forms in this example or in the experiment below, so I stick with the simpler interpretation of harmony scores.

- Ratings are plausibly influenced by many different factors: grammatical well-formedness, similarity to existing words, ability to assign meaning, random noise
    - ○ Is difference between [mlit] and [mlɔʒ] due to a difference in grammatical well-formedness, or to greater similarity to existing words (e.g., *mit*, *flit*, *Brit*, *quit*)?
    - ○ Similarly, is preference for [mɹuˈpeɪʃən] due to phonotactic likelihood of [ˈpeɪʃən], or to the fact that it contains recognizable English morphemes (*-ation*), which somehow trumps phonotactics?

3 The data

(10) Acceptability ratings of 210 nonce monosyllables

- Items embodied a broad and continuous range of expected values, as predicted by joint conditional bigram probability and by support in Generalized Neighborhood Model (Bailey and Hahn 2001), based on counts from CELEX (Baayen, Piepenbrock, and van Rijn 1993)
    - ○ Well supported: [krʌsp], [pʌm], [smæt], [mɪm], [paɪt]
    - ○ Intermediate support: [miːv], [fræg], [ʃrʌt], [wɛʃt], [snʌlk]
    - ○ Low support: [tlad], [pniːk], [θnɔf], [ʒnɛt],
- Recorded by a native speaker of English in simple carrier frames, as nouns or verbs
    - ○ [θnɔf]. This is a [θnɔf].
    - ○ [θnɔf]. I like to [θnɔf].
- Stimuli checked by two phonetically trained listeners to ensure presence of all acoustic landmarks for all consonants in clusters (Stevens 1998; Stevens 2002)
- Presented to subjects auditorily, in random order
    - ○ Noun and verb presentations counterbalanced across subjects; each subject rated equal numbers of noun and verb frames
- Subjects repeated word, and rated on scale of 1 (=worst) to 7 (=best)
    - ○ Repetitions transcribed by two phonetically trained listeners; rating from trials with incorrect repetition were discarded
    - ○ No significant difference between noun and verb ratings (Albright, in press); ratings combined for present purposes

(11) Strategy

- Full set of 205 items included items with a wide variety of different kinds of structures (rare/illegal onset clusters, codas, VC combinations, etc.)
- More controlled comparison: select batches of items which are similar in one respect (i.e., contain a rare or illegal cluster) and differ in another respect (probability of remainder of the word)
- In the following sections, we analyze subsets of this larger set, selected to test specifically for cumulative effects

4 Cumulative evaluation of onsets and rhymes

Question 1: are outputs as bad as their worst violation?

- Are items that share the same illegal cluster rated, on average, as equally unacceptable, since they are eliminated by the same constraint?
- If there are significant differences in their ratings (as in [mlit], [mlɔʒ] example above), are these differences explained by the likelihood of their rhymes?
- Is the influence of rhymes best modeled with the grammar, or an independent effect of lexical similarity?

(12) Test items

- Selected 39 nonce words with unattested onset clusters, paired with varied (but legal) rhymes
  - ○ *#pw*[2], *#bw, #tl, #dl*; *#fn, #θn, #pn, #bn*; *#pt, #bd, #bz*

    pwæd
    pwɛt
    pwɪst
    pwʌdz
    pwʌs
    tleɪk
    tliːn
    tlad

    bwæd
    bwad
    bwʌd
    dliːk
    dliːn
    dlaɪk
    dlʌd
    fnɛdʒ

    fnɪtʃ
    fnoʊ
    fnuːt
    θnɛdʒ
    θnuː
    θnɔf
    pniːk
    pniːn

    pnɛp
    bniːn
    bnad
    bnʌs
    ptæd
    ptiːn
    ptɛp
    ptʌs

    bdiːk
    bdʌs
    bduːt
    bziːn
    bzaɪk
    bzad
    bzʌs

- Preliminary indication that multiple violations do, in fact matter: acceptability ratings of words sharing the same illegal onset cluster (1 = worst, 7 = best)

bziːn	2.00	ptiːn	2.44
bzʌs	1.81	ptʌs	1.94
bzad	1.63	ptɛp	1.86
bzaɪk	1.28	ptæd	1.67

  - ○ Considerable differences among sets of items, even though all ruled out by the same constraints on onset clusters (e.g., *#[−strid][−son])
  - ○ Plausibly related to differences in likelihood/well-formedness of rhymes
  - ○ Mirrors Ohala and Ohala (1986) result, but in this case the second violation (the VC combination) would not normally be fatal in English

- Strategy: compare models based on different modes of evaluation
  - ○ Grammatical well-formedness based on worst violation (here, the onset cluster)
  - ○ Cumulative grammatical well-formedness across the word
  - ○ Combination of grammatical well-formedness and analogy to existing words

[2] I include *#pw*/*#bw* as unattested, although they occur (and are pronounced faithfully) in familiar loanwords such as *pueblo* and *bueno*. Moreton (2002) provides evidence that these clusters are not dispreferred to the same extent as truly unfamiliar clusters (such as *#dl*), though they are certainly marginal compared to clusters such as *#pr, #br*. Ultimately, the classification as "attested" or "unattested" is not crucial to the analysis below, which assigns each cluster a numerical well-formedness score. All that matters for present purposes is that the initial cluster is worse than any of the rhymes it is paired with, and should thus produce the highest ranked violation.

4.1 Modeling gradient grammatical well-formedness based on lexical frequency

(13) A starting assumption

- Gradient well-formedness of a structure is related to its rate of attestation in the lexicon (Coleman & Pierrehumbert 1997; Frisch, Large, & Pisoni 2000; Hay, Pierrehumbert, & Beckman 2004; Vitevitch & Luce 2004; Hayes and Wilson, in press)
- All onset clusters under consideration here are unattested in English (prob = 0)
- However, when viewed in terms of features/natural classes, some clusters represent more widely attested combinations than others
  - ○ *#pw*, *#tl*: stop+glide, stop+liquid (well attested)
  - ○ *#fn*, *#θn*: fricative+nasal (moderately well attested)
  - ○ *#pn*, *#bn*: stop+sonorant (attested, at least in form of stop+glide and stop+liquid)
  - ○ *#bd*, *#bz*: stop+obstruent: unattested

(14) An inductive model of gradient well-formedness (Albright, in prep.)

- Each sequence that is attested in the training corpus provides evidence not only about that particular combination of segments, but also about combinations of natural classes
- Intuition: sequences provide strong support for the specific combinations that they embody, weaker support for more general descriptions
- Example: *play*, *plow*, *plant*, etc. provide...
  - ○ Strong support for the legality of *#pl*
  - ○ Somewhat weaker support for the legality of *#p*+liquid, #labial+*l*, etc.
  - ○ A little bit of support for the legality of #stop+liquid
  - ○ A tiny bit of support for the legality of #stop+sonorant
  - ○ Practically no support for the legality of #stop+consonant combinations
- Goal: given a novel sequence, find that combination of natural classes that has received the greatest support from the training corpus

(15) Formalization

a. Likelihood of sequence *ab*, where $a \in$ class A, $b \in$ class B

$$\propto \frac{\text{\# of times } ab \text{ occurs in corpus}}{\text{Total \# of two-item sequences}} \times \text{Prob(choosing } a \text{ from } A) \times \text{Prob(choosing } b \text{ from } B)$$

b. $\text{Prob(choosing } a \text{ from class } A) = \dfrac{\text{Count of } a \text{ in the corpus}}{\text{Count of all members of } A}$

c. Given sequence *ab* to evaluate, find those classes *A*, *B* that maximize the likelihood of the sequence

☞ For details, see http://www.mit.edu/~albright/papers/Albright-GrammaticalGradience.pdf

(16) Consequence

- For sequences that co-occur sufficiently often, this favors characterizations in terms of specific combinations of classes
  - ○ $P(br) = P(\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{labial} \end{bmatrix} [+\text{rhotic}]) \times P(b \mid \text{labial stops})$ (quite high)

- For sequences that rarely or never co-occur, it is necessary to seek more general characterizations, at a cost (denominator in (15b))
  - ○ $\mathrm{P}(bw) = \mathrm{P}(\begin{bmatrix} +\text{lab} \\ +\text{voi} \end{bmatrix} [-\text{cons}]) \times \mathrm{P}(b \mid \text{vcd labials}) \times \mathrm{P}(w \mid \text{vowels/glides})$
  - ○ $\mathrm{P}(bd) = \mathrm{P}(\begin{bmatrix} -\text{cont} \\ +\text{voi} \end{bmatrix} \begin{bmatrix} -\text{cont} \\ -\text{son} \\ +\text{cor} \\ +\text{voi} \end{bmatrix}) \times \mathrm{P}(b \mid \text{vcd stops/nasals}) \times \mathrm{P}(d \mid \{d, dʒ\})$

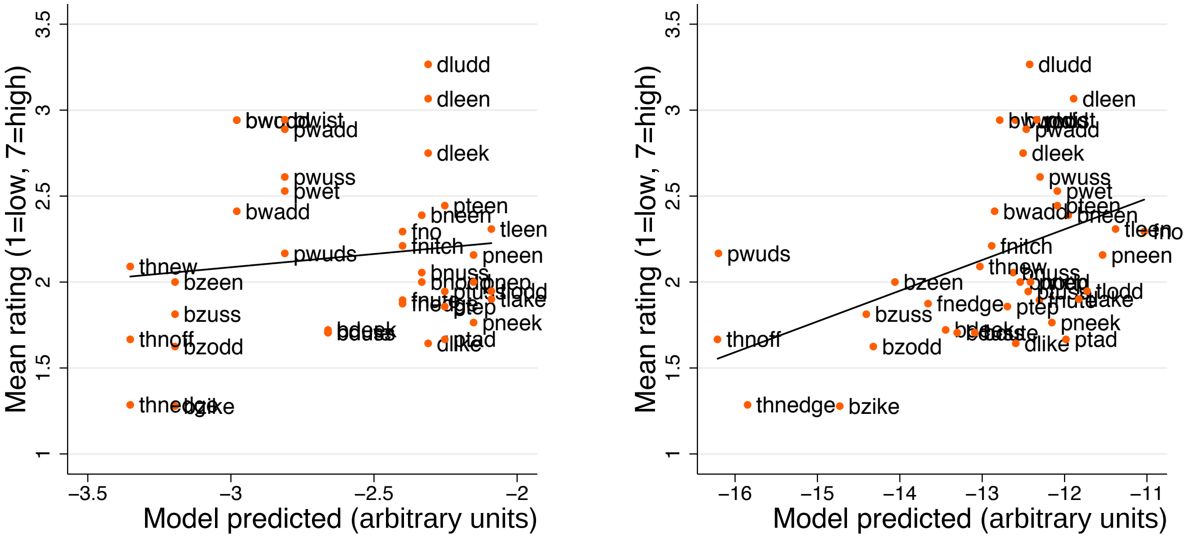
(17) Modeling ratings of nonce words

- Recall the first possibility, inspired by OT
  - ○ Acceptability of a word is determined by its highest ranked violation (i.e., well-formedness of the worst subsequence, which for these items is the onset cluster)
  - ○ Model: for each nonce word, assign a score equal to probability of its onset cluster, as defined above
- Alternative possibility: cumulative evaluation
  - ○ Acceptability of item is determined jointly by well-formedness of all subparts
  - ○ Model: for each nonce word, calculated the summed log probability of all subsequences (i.e., the log of the product of the probabilities)

(18) Results

a. Onsets only (r(37) = .13)

b. Whole words (r(37) = .45)



- Cumulative model is clearly better (graph (b.), on right), suggesting rhymes do influence ratings in an interpretable way
- Both models leave substantial variability unaccounted for, even concerning onsets
- Notably: overestimate acceptability of sequences like *#pn*, *#pt*, and underestimate acceptability of sequences like *#bw*, *#dl* relative to other clusters
- Suggests that this way of counting statistics over natural classes is not sufficient to capture relative well-formedness of unattested onset clusters

☞ A role for markedness?

4.2 Incorporating markedness biases

(19) The clusters in (12) show several possibly distinct markedness effects

- Preference for obstruents to occur before segments that support perception of burst and formant transitions → strongly voiced C_2 (≈ sonority sequencing)
  - ○ C_2 SONORITY: $*\begin{bmatrix}-\text{continuant}\\-\text{sonorant}\end{bmatrix}$ /___ clear formant structure
  - ○ Assumption: violations reflect availability of formant structure[3]

	C_2 SON
stop+glide	*
stop+liquid	**
stop+nasal	******
stop+obstruent	*******

- Avoidance of sequences in which C_2 obscures transitions of C_1 by having similar targets → different places of articulation (≈ OCP)
  - ○ OCP labial: violated by *#pw*, *#bw*
  - ○ OCP coronal: violated by *#tl*, *#dl*

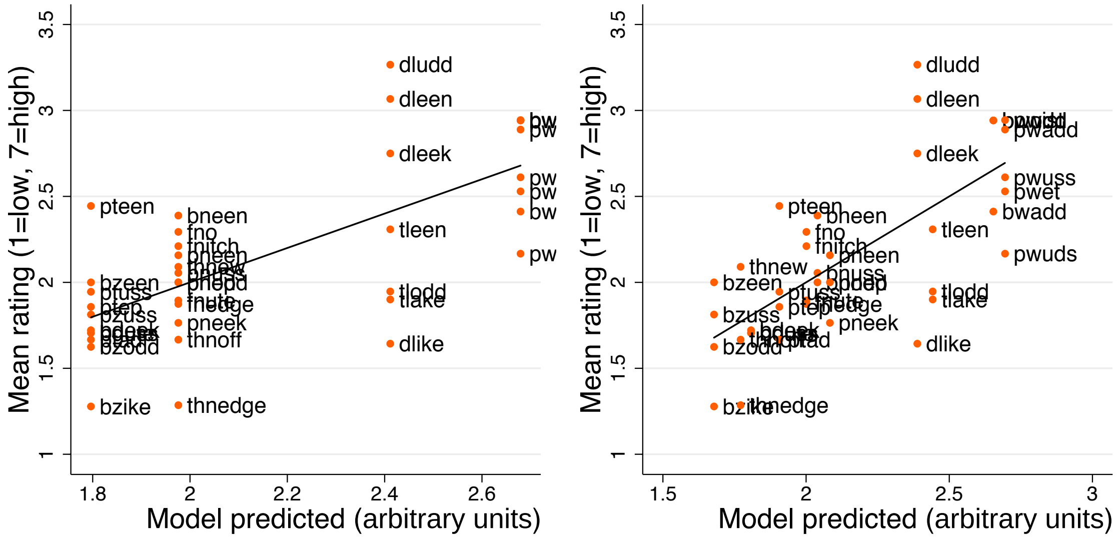
(20) Augmenting the statistical model

- To give a fair test of cumulative evaluation across the entire word, we want predictions about the well-formedness of onset clusters that are as accurate as possible
- Heuristic strategy: add markedness constraints to the model, and determine their numerical importance using a Generalized Linear Model, fitted to the experimental ratings (Baayen, in press)[4]
- As before, compare results based on evaluation of onset clusters, or of entire word

(21) Results

a. Markedness constraints alone (r(37) = .695)

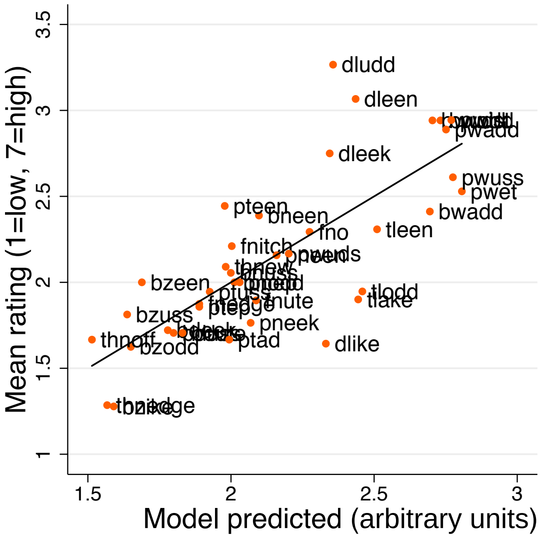
b. Markedness + onset probability (r(37) = .717)



[3] For justification, see: http://www.mit.edu/~albright/papers/Albright-BiasedGeneralization.pdf

[4] Keller (2006) discusses the relation between Generalized Linear Models and grammars of weighted constraints.

c. Markedness + statistical evaluation of entire words (r(37) = .772)



- Substantial improvement over models based on statistics alone
- Some onset effects still not captured (e.g., relative preference for *#dl* as compared with *#tl*), but reasonably good fit overall
- Model that assesses probability across the entire word is still most accurate (in (c.))
- Effect is subtle, since rhymes here were chosen to all be relatively well-formed (to guarantee that onset violation is worst)
- In section 5, we examine the simultaneous effect of somewhat worse violations in both onset and rhyme

(22) Conclusion so far

- Best "purely grammatical" model of acceptability ratings is one that employs combination of markedness biases and statistical evaluation over the entire word
- Model based on statistics alone, and model based on markedness + statistical probability of the worst violation, are not as accurate
- Appears to support a cumulative model of evaluation
- However, leaves open other possibilities
  - ○ Perhaps cumulative effects could emerge through standard OT-style evaluation, but applied comparatively across output candidates (Coetzee 2004)
  - ○ Influence of the rhyme is not a grammatical effect at all, but rather, the influence of a separate mechanism: analogical support by lexical neighbors

4.3 Comparative evaluation of outputs within OT

(23) Assumption thus far

- Output forms that are eliminated by the same fatal violation should receive the same grammatical well-formedness score
- Violations of lower-ranked constraints are irrelevant

(24) Another possibility: comparative evaluation of outputs for different inputs (Berent, Everett, and Shimron 2001; Coetzee 2004)

- Simple case: comparison of outputs that differ on ranking of highest violation

	*[bn	Dep	*lθ]	*[sf
/bnɪk/ → bnɪk	*!			
/sfɪk/ → sfɪk				*

sfɪk ≻ bnɪk

- In side-by-side comparison, [sfɪk] is more harmonic since worst violation is of a lower ranked constraint
- However, side-by-side evaluation is not really needed; all we need access to is rank of highest violation of form under consideration, and some global metric relating this rank to ratings

(25) A less clear case: outputs under consideration share the same worst violation

	*[bn	Dep	*lθ]	*[sf
/bnɪk/ → bnɪk	*!			
/bnælθ/ → bnælθ	*!		*	

bnɪk ≻ bnælθ ?

- Unlike above, these outputs are not differentiated by rank of highest violation
- Under standard OT evaluation, [bnɪk] would indeed be more harmonic than [bnælθ] if treated as competitors for same input (mark cancellation), so it is conceivable that comparative evaluation could yield [bnɪk] ≻ [bnælθ]
- However, experimental task being modeled here is not explicitly comparative—e.g., [bniːn] on trial 37, [bnad] on trial 102
- During individual trials, subjects presumably have limited ability to compare current output to similar outputs
    - ○ More plausibly comparing nonce word to competing candidates for same input
    - ○ For this comparison, violations below the fatal violation are irrelevant—and in fact, would normally be ignored by standard OT evaluation procedures, which discard candidates as soon as they are eliminated
- It is hard to see why non-fatal violations of output forms like [bnælθ] would still be accessible on subsequent trials involving other *#bn*-initial items

(26) Provisional conclusion

- Not clear that cumulative effects seen here can be appropriately modeled using comparative evaluation in a non-cumulative model (OT)
- For related arguments in favor of handling cumulativity effects with an additive model, see Pater (2007)

4.4 The analogical influence of the lexicon

(27) Modeling lexical support (analogy)

- A traditional measure: number of neighbors
    - ○ Number of existing words that differ by changing, adding, or deleting a single segment (Greenberg and Jenkins 1964; Coltheart, Davelaar, Jonasson, and Besner 1977; Luce 1986)

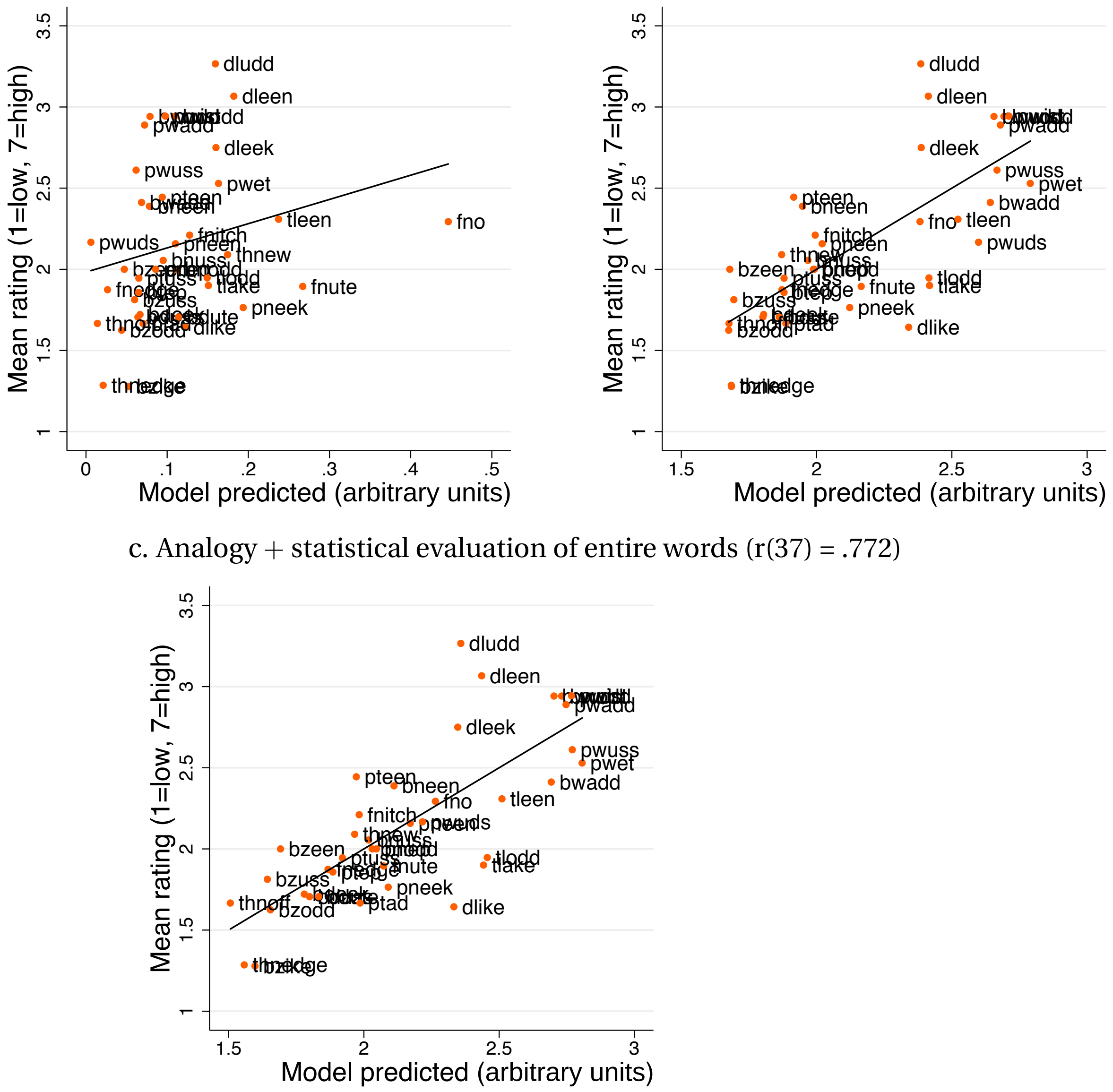
- Bailey and Hahn (2001): Generalized Neighborhood Model (GNM)
    - ○ Lexical support = similarity to existing words, summed over the entire lexicon
    - ○ Probability(novel word) $\propto \sum$ Similarity(novel word,existing words)
    - ○ Similarity to existing words determined by a string alignment algorithm, sensitive to phonetic distance between corresponding segments
    - ○ To be well supported by the lexicon, a novel word should be relatively similar to a decent number of existing words
    - ○ For details, see Bailey and Hahn (2001); Albright and Hayes (2003)

(28) Testing the influence of analogy on acceptability ratings

- Used implementation of Bailey and Hahn's GNM to estimate degree of lexical support for experimental test items
- Combined grammatical and analogical predictions using Generalized Linear Model

(29) Results

a. Analogy alone (r(37) = .244) b. Analogy + onset probability (r(37) = .737)



c. Analogy + statistical evaluation of entire words (r(37) = .772)

- Lexical analogy alone is not a decent predictor of experimental ratings
- Combining analogical and onset effects does yield a small improvement over onset-based probability alone, but improvement is very slight
- For additional arguments that analogy appears to play very little role in experiments involving ratings of nonce words, see Shademan (2007), Albright (in prep.)
- Best model incorporates probabilities of all sequences within the word

(30) Local summary

- Words that share the same fatal violations do receive substantially different acceptability ratings depending on the material in the rest of the word
- When various models are considered alone and in combination, the most effect model is also one of the simplest: acceptability ratings are the result of evaluating likelihood of co-occurrence over the entire word
  - ○ Likelihood determined partly by probability within the lexicon, and partly by markedness biases
- Supports a cumulative model, such as the inductive model employed here or a linear/additive model of constraint interaction

5 Ganging up effects

Question 2: can constraint violations gang up to overcome violations of higher constraints?

- Do violations of higher-ranked constraints strictly outweigh violations of lower-ranked constraints?
- Preliminary support for ganging up in nonce word ratings, from Coleman and Pierrehumbert (1997): [mɹuˈpeɪʃən] (one illegal cluster) > [ˈsplɛtɪsak] (several low-probability structures)
- Can similar effects be observed in more controlled comparisons?

(31) Strategy

- We seek to test whether multiple violations of weaker constraints can *in principle* gang up to overcome a stronger constraint
- This requires a very particular type of configuration
  - ○ weight(weak constraint$_1$) + weight(weak constraint$_2$) > weight(stronger constraint)
- Nothing guarantees that configuration will hold; in fact, strongly enforced constraints may easily have weights > twice the weights of weaker constraints
- Strategy: focus on "weak phonotactic effects", inviolable in the language, but near the cut-off of a scale, parochial (not part of a much more general phonotactic effect)

(32) Test items

a. Group A: 20 items with one mild but unattested phonotactic violation

- **#bw*: bwɑd, bwʌd, bwæd
- **#dl*: dlʌd, dliːn, dliːk, dlaɪk

- **s*NVN (Davis 1984): snʌmp, smʌm, smɪmp
- *nasal+non-coronal voiced stop/___ #: tɛmb, fɹæmb, hɪmb, dʒæŋg, vɪŋg
- **stw* (Clements and Keyser 1983): stwɪp, stwaɪm
- Other unattested codas: wɛʃt, gɹɛʃt, ɹɪnv

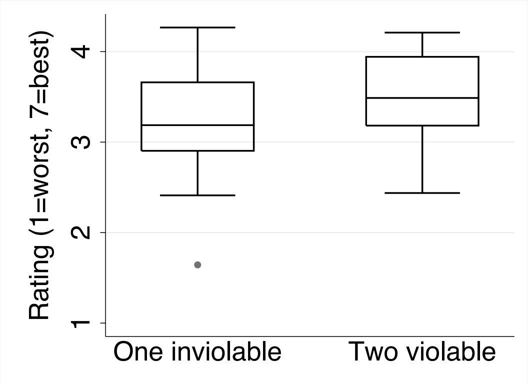
b. Group B: 20 items with two rare(-ish) but attested margins

zɪndʒ (cf. *zinc, twinge*)
snʌlk (cf. *snuck, skulk*)
pɹʌpt (cf. *prim, erupt*)
fɹɛkt (cf. *friend, effect*)
θɹɛlt (cf. *threat, melt*)
gɹɑlf (cf. *graph, golf*)
tɹɪlb (cf. *trick, bulb*)
ʃɛsp (cf. *shed, asp*)
bɹɛlθ (cf. *bread, health*)
bɹɛnθ (cf. *bread, plinth*)

dwoʊdʒ (cf. *dwell, doge*)
fɹɛg (cf. *friend, egg*)
glæg (cf. *glass, bag*)
glækt (cf. *glass, act*)
niːps (cf. *kneel, traipse*)
twiːks (cf. *twin, hoax*)
gwɛpt (cf. *Gwen, wept*)
spɛlʃ (cf. *spell, welsh*)
sklaɪm (cf. *sclerosis, slime*)
skluːnd (cf. *sclerosis, wound*)

(33) Preliminary observation: substantial overlap in ratings

- No significant preference for “legal” items (F(1,38) = 1.99, p = .17)

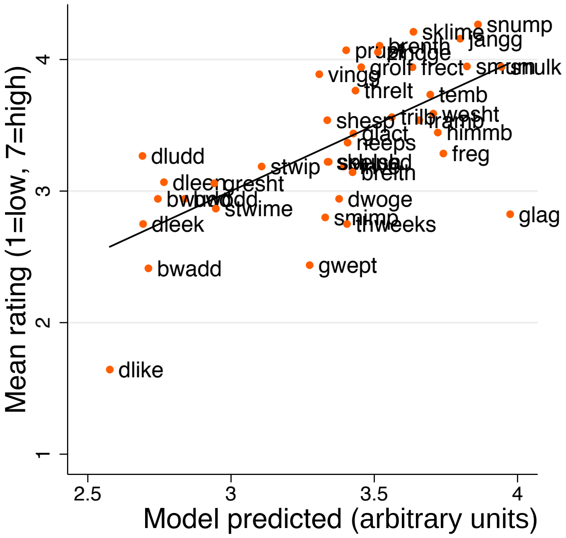


- Examples
  - snʌmp, smʌm > fɹɛkt, gɹɑlf, θɹɛlt
  - tɛmb, fɹæmb, hɪmb > glækt, niːps, bɹɛlθ
  - dlʌd, dliːn > dwoʊdʒ, gwɛpt
- This simply replicates Coleman and Pierrehumbert’s [mɹuˈpeɪʃən] > [ˈsplɛtɪsak] observation, but in shorter words with fewer potential confounds

(34) Modeling these results

- As above, used Generalized Linear Model with cumulative predictions of natural class-based inductive model for the entire word, together with manually entered markedness constraints
  - C2 sonority, OCP-Lab, OCP-Cor, *nas+noncoronal stop#
  - *ʃC, *s[+nas], *stw

- Results: a reasonable fit to the data (r(38) = .656)



- Crucially, no clear separation between acceptability ratings of items with one illegal structure and those with two legal but rare structures
- Favors a model that allows ganging up, as in Harmonic Grammar, maxent models, or the inductive model employed here

(35) Acceptability vs. grammaticality

- Although ratings do not cleanly distinguish items normally classified as grammatical in English (such as [θɹɛlt] or [glækt]) from those that are not ([dliːn], [tɛmb]), this does not mean that grammaticality is a figment or irrelevant
- An important distinction: (Schütze 2005)
  - Acceptability: a surface evaluation of well-formedness of surface form, as fully faithful candidate of corresponding input (e.g., /bnɪk/ → [bnɪk])
  - Grammaticality: a comparative evaluation, the probability that surface form would actually be chosen as the surface form of the corresponding input[5]
- Schematic example:

/tɛmb/	→	[tɛmb]	−2.5
		[tɛm]	−2.0
/glækt/	→	[glækt]	−3.0
		[glæk]	−3.5

  - Output [tɛmb] is more harmonic/acceptable than [glækt]
  - However, [tɛmb] loses to [tɛm] in grammatical competition (ungrammatical), while [glækt] wins (grammatical)

[5] Pater (2007) uses the term acceptability for what I am calling here grammaticality. Pater's usage is designed to avoid having ungrammatical items receive higher acceptability scores than ungrammatical items; however, that is precisely what we need in the present case.

6 Discussion and conclusion

(36) Results here support two claims made previously in the literature based on smaller numbers of forms, or forms not designed to test these issues specifically

- Acceptability ratings of nonce forms depend on features of the entire word, and not solely on the severity of the worst violation
- Items with two low probability sequences can be deemed worse than items with one unattested (arguably ungrammatical) sequence

(37) These results are a challenge to standard interpretations of OT, and support a model that combines gradient well-formedness values from across the entire word

- Probabilistic model of parsing into constituents (Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000; Hay, Pierrehumbert, and Beckman 2004)
- Additive model using weighted constraints (Pater 2007; Hayes and Wilson, in press)
- N-gram style model based on natural classes (Albright, in prep.)

(38) Although neither of these claims is new, the results here do sharpen the issue somewhat

- Experimental test with larger number of items in a more constrained domain show the effects more systematically than in previous studies
- Comparison with partly analogical model shows that these effects cannot be written off as the independent influence of similar existing words

(39) Why aren't cumulative violations ever categorically fatal cross-linguistically?

- Under the interpretation employed here (and according to experimental ratings), [bnælθ] is deemed worse than [bnɪk] by the grammar
- A reasonable expectation: languages might draw a cut-off between these forms, banning doubly marked [bnælθ] but allowing the simpler form [bnɪk]
- This appears not to happen cross-linguistically; why not?

(40) One possibility (Pater, Bhatt, and Potts 2007)

			*[bn	Dep	*lθ]	*[sf
	a.	bnɪk	*!			
☞	a.′	bənɪk		*		
	b.	bnælθ	*!		*	
☞	b.′	bənælθ		*	*	
	b.″	bənæləθ		**!		

- In order for a language to ban doubly-marked [bnælθ], the constraints must be weighted in such a way that the grammar favors some particular repair (e.g., /bnælθ/ → [bənælθ])
- However, in order for the grammar to prefer [bənælθ] over [bnælθ], the following condition must hold: weight(*#bn) > weight(*Dep*)
- If this condition holds, then epenthesis will be preferred for any input with /#bn/, regardless of what violations occur in the rest of the word
- Upshot: lack of grammaticized cumulativity effects is a consequence of the way that markedness and faithfulness interact to define possible outputs

(41) Another possibility: perhaps grammars *can* ban doubly-marked forms

- Bans on doubly marked structures do seem to occur as stages during acquisition
    - E.g., children may go through a stage in which complex onsets and complex codas are allowed, but not simultaneously within the same word (Levelt, Schiller, and Levelt 2000; Boersma and Levelt 2000)
- Doubly marked forms may also be statistically underrepresented in adult language
    - Counts over 6292 monosyllabic lemmas in CELEX reveal that with singleton onsets, *s#* and *s*+stop# codas occur in comparable numbers (with slight preference for coda clusters, even)
    - However, with less common stop+liquid onset clusters, coda /sC/ appears to be relatively underrepresented

Onset	s	sC
bV___	10	17
dV___	7	5
gV___	8	6
pV___	15	11
tV___	3	9
lV___	12	12
rV___	3	16

Onset	s	sC
grV___	11	3
trV___	7	3
plV___	6	1
drV___	3	0
glV___	3	0

-
    - Conversely, /lC/ codas are underrepresented when they occur with /sC/ onsets

Onset	l	lC
bV___	23	20
gV___	13	11
wV___	25	25
mV___	16	17
sV___	16	14
pV___	27	9
tV___	18	6

Onset	l	lC
spV___	8	1
stV___	17	3
swV___	7	0
snV___	3	0

- Albright, Magri and Michaels (in press)
    - In order for grammar to allow single marked structures but ban combinations of marked structures, a very specific condition must hold:

      2×weight(markedness) > weight(faithfulness) > weight(markedness)
      (*doubly marked structures banned*) (*singly marked structures allowed*)

    - If training on singly marked structures pushes faithfulness and markedness far apart (weight($\mathcal{F}$) $\gg$ weight($\mathcal{M}$)), then multiple markedness violations can no longer gang up to overcome faithfulness
    - Learning bias for constraint weights to be far apart explains why such stages are seen during acquisition or as gradient preferences, but not as categorical bans

(42) On-going efforts

- Integrating phonetically motivated markedness biases and statistically learned constraints into a unified learning model
- Investigating evidence for cumulative effects not only in ratings, but also in gradient avoidance of doubly marked structures (and as an acquisition stage)
- Reconciling these effects with lack of categorical ganging up in adult languages

References

Albright, A. (in prep.). Gradient phonological acceptability as a grammatical effect. MIT ms.

Albright, A. (in press). How many grammars am I holding up? Discovering phonological differences between word classes. In *WCCFL 26*. Cascadilla Press.

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

Albright, A., G. Magri, and J. Michaels (in press). Modeling doubly marked lags with a split additive model. In H. Chan, H. Jacob, and E. Kapia (Eds.), *BUCLD 32: Proceedings of the 32nd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Baayen, R. H. (in press). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

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.

Berent, I., D. Everett, and J. Shimron (2001). Do phonological representations specify variables? evidence from the obligatory contour principle. *Cognitive Psychology 42*(1), 1–60.

Boersma, P. (2004). A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments. Unpublished ms.

Boersma, P. and C. Levelt (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of Child Language Research Forum*, Volume 30, Stanford, pp. 229–237. CSLI Publications.

Chomsky, N. and M. Halle (1965). Some controversial questions in phonological theory. *Journal of Linguistics 1*, 97–138.

Clements, G. N. and S. J. Keyser (1983). *CV Phonology*. Cambridge, MA: MIT Press.

Coetzee, A. W. (2004). *What it Means to be a Loser: Non-Optimal Candidates in Optimality Theory*. Ph. D. thesis, University of Massachusetts, Amherst.

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

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

Davis, S. M. (1984). Some implications of onset-coda constraints for syllable phonology. In *Chicago Linguistic Society*, Volume 20, pp. 46–51.

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.

Frisch, S. A. and B. A. Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language 77*, 91–106.

Goldwater, S. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory, Stockholm University*.

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

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

Jäger, G. (to appear). Maximum entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (Eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford: CSLI Publications.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of gramaticality*. Ph. D. thesis, Univ. of Edinburgh.

Keller, F. (2006). Linear Optimality Theory as a model of gradience in grammar. In G. Fanselow, C. Féry, R. Vogel, and M. Schlesewsky (Eds.), *Gradience in Grammar: Generative Perspectives*, pp. 270–287. Oxford University Press.

Kirby, J. and A. Yu (2007). Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In *Proceedings of the International Congress of the Phonetic Sciences XVI.*, pp. 1389–1392.

Levelt, C., N. O. Schiller, and W. J. Levelt (2000). The acquisition of syllable types. *Language Acquisition 8*, 237–264.

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

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

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.

Pater, J. (2007). Cumulative ill-formedness in typological and experimental data. Paper presented at the Conference on Experimental Approaches to Optimality Theory, University of Michigan, May 2007.

Pater, J., R. Bhatt, and C. Potts (2007). Linguistic optimization. University of Massachusetts, Amherst ms.

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

Shademan, S. (2007). *Grammar and Analogy in Phonotactic Well-formedness Judgments*. Ph. D. thesis, University of California, Los Angeles.

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America 111*, 1872–1891.

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

Vitevitch, M. 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.