Similarity, feature-based generalization and bias in novel onset clusters

Adam Albright

Massachusetts Institute of Technology

8 July 2007

Introduction

Well-established relation between lexical statistics and gradient acceptability

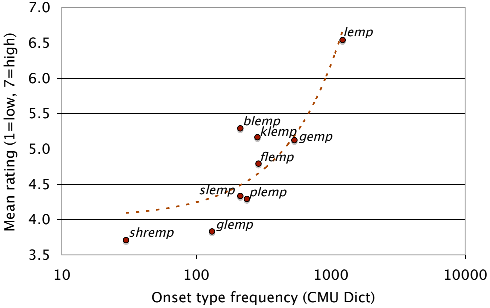
- Novel items with high-frequency combinations of phonemes, morphemes, etc., tend to sound more "English-like" than items with rare or unattested combinations

E.g., for phonotactics:

Coleman and Pierrehumbert (1997); Treiman, Kessler, Knewasser, Tincoff, and Bowman (2000); Frisch, Large, and Pisoni (2000); Bailey and Hahn (2001); Hammond (2004); Hayes and Wilson (in press); and many others

Example: novel words ending in __ *emp*

- Bailey and Hahn (2001): wordlikeness ratings of novel words ("How typical sounding is *blemp*?")
- Correlated against type frequency of onsets from CMU Pronouncing Dictionary, counted by Hayes & Wilson (in press)
- Clear preference for more frequently attested onsets



The limits of attestedness

- Growing body of literature investigating preferences that do not follow straightforwardly from statistics of the input data
  - ☞ Preference for some attested sequences over others (Moreton 2002, 2007; Wilson 2003; Zhang and Lai 2006)
  - ☞ Preference for some unattested sequences over others (Wilson 2006; Finley & Badecker 2007; Berent & al., in press)
- Such cases have potential to reveal substantive analytical bias (Wilson 2006)

Example

Berent et al. (in press)

- English speakers prefer initial *#bn* over *#bd*
  - More likely to interpret [bdɪf] as [bədɪf], without a cluster
- Little direct evidence in favor of *#bn* ≻ *#bd*
  - Few if any attested examples: *bnai brith*, *bdellium*
  - Very few words that could potentially exhibit initial /bən/, /bəd/ → [bn], [bd] (*beneath*)
  - Also few words with medial /bən/, /bəd/ → [bn], [bd], (*nobody*, *ebony*, *Lebanon*; generally fail to syncopate)
  - In final position, [bd] is well attested (*grabbed*, *described*), but [bn] is unattested
- Preference evidently not due to greater exposure to [bn]
  - Perhaps due to bias towards rising sonority profiles?

Indirect generalization

- Although English speakers have relatively little direct experience with [bd], [bn], they have plenty of experience with clusters like [bl] and [br]
- More generally, initial stops are always followed by a sonorant (C_2 or vowel)
- Perhaps preference for *#bn* could be inferred from distribution of occurring clusters

Perceptual similarity

☞ *#bn* perceptually closer to *#bl* than *#bd* is (?)

Featural similarity

☞ *#bn* part of a broader pattern of stop+coronal sonorant sequences (Hammond, Pater yesterday)

Goals of this talk

- Report on some attempts to model preferences like $\#bn \succ \#bd$ based on indirect inference from attested clusters
  - Test the extent to which they can be predicted by a statistical model, without prior markedness biases
  - Of course, a successful data-driven model doesn't *prove* that humans learn similarly
  - However, the case for prior bias is diminished
- Preview: mixed results
  - Some preferences potentially learnable, given certain assumptions (e.g., $\#bn \succ \#bd$)
  - Others not learned by any model tested so far ($\#bw \succ \#bn$)
- Provisional claim: best model of speaker preferences combines learned statistical generalizations and markedness biases

Outline

- Compare two models of gradient acceptability of attested sequences
    - A feature-based grammatical model
    - A similarity-based analogical model (Generalized Neighborhood Model; Bailey and Hahn 2001)
- Test models' ability to capture preferences among unattested onset clusters, by generalization from attested clusters
    - Sonority preferences in stop+C clusters
    - Sonority + place preferences in #*bw* vs. #*dl*
    - Place preferences in *s*C clusters
- Pay-off of combining phonetic biases with learned statistical preferences

What we want a model to do

Some desiderata for a statistical model of gradient phonotactics

- Trained with realistic L1 input
  - Child-directed data
  - Approximated here with adult corpus data (CELEX)
- Predict relative preferences among combinations of attested sequences
  - *#kl*, *#fl* $\succ$ *#gl*, *#sl*
- Predict relative preferences among combinations of unattested sequences
  - *#bw*, *#bn* > *#bd*, *#bz*
- Able to make predictions for entire words
  - *stip* [stɪp], *plake* [pleɪk] $\succ$ *chool* [tʃuːl], *nung* [nʌŋ]
  - *mrung* [mrʌŋ], *vlerk* [vlr̩k] $\succ$ *shpale* [ʃpeɪl], *zhnet* [ʒnɛt]

Two types of generalization

Two fundamentally different modes of generalization

- Comparison to the lexicon: how similar are *blick*, *bnick* to the set of existing words?
  - ≈ 'Dictionary' task (Schütze 2005)
- Evaluation of substrings: how probable/legal are the substrings that make up *blick*, *bnick*?
  - ≈ Grammatical acceptability
- Plausible that speakers perform both types of comparison, to varying degrees depending on the task (Vitevitch & Luce 1999; Bailey & Hahn 2001; Shademan 2007; and others)

Goal of this section

- Sketch models that instantiate these two types of generalization
- Present benchmarking results on two types of data
  - Ratings of nonce words with (mostly) attested sequences
  - Ratings of a mix of attested and unattested onset clusters (Scholes 1966)

☞ Although neither model is perfect, both provide a reasonable first-pass estimate of gradient phonotactic acceptability for these data sets

Neighborhood density

A crude but widely used estimate of similarity to the lexicon: neighborhood density

- Number of words that differ from target word by one change (Greenberg & Jenkins 1964; Coltheart, Davelaar, Jonasson & Besner 1977; Luce 1986)
- Generally inadequate for non-words: most have few or no one-change neighbors (Bailey and Hahn 2001)

The Generalized Neighborhood Model

Bailey and Hahn's (2001) Generalized Neighborhood Model

- Support depends on gradient similarity to existing words, rather than one-change threshold
- Prob(novel word) $\propto \sum$ Similarity(novel word,existing words)
- Every existing word contributes some support, but in most cases it's quite small
- To be well supported by the lexicon, a novel word should be relatively similar to a decent number of existing words (for model details, see Bailey and Hahn 2001)

The Generalized Neighborhood Model

Model parameters

- Lexicon modeled as set of lemmas with freq > 0 in CELEX
- Similarities calculated using natural class model of Frisch, Pierrehumbert and Broe (2004)
- No advantage found for using surface word forms, or token frequency
- Remaining parameters selected by fitting

Testing the model

Benchmarking data

- 92 wug words, used in pre-test to past tense study (Albright and Hayes 2003)
    - A few rare or illegal sequences (*#ʃw*, *Vːnθ#*, etc.)
- 70 wug words with no unattested sequences (Albright, in prep.)
    - Chosen randomly from set of 205 words used in a larger study

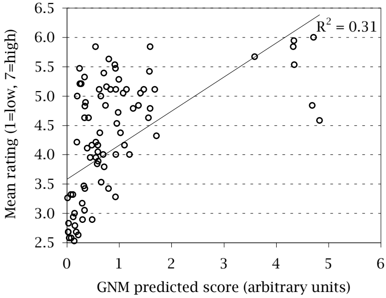
Testing the model

Experimental details

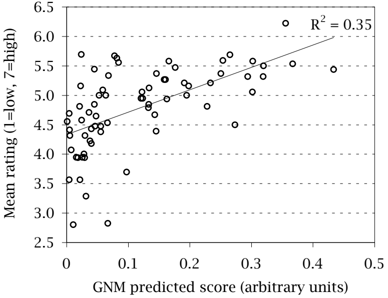
- Presented auditorily in simple frame sentences (e.g.: '[stɪp]. *I like to* [stɪp].')
- Subjects repeated aloud, and rated from 1 (implausible as an English word) to 7 (would make a fine English word)
- Ratings discarded from trials in which subjects repeated word incorrectly

GNM results

Albright and Hayes (2003) data: ($r(92) = .557$)



Albright (in prep.) data: ($r(70) = .592$)



- Significant moderate fit, though room for improvement
- Possibly improved by non-linear fit (Hayes and Wilson, to appear)
- Reasonable first pass estimate of gradient acceptability

Biphone probabilities

Simple model of attested sequences: biphone probabilities

- Biphone = sequence of two segments
- Probability of a two-segment sequence:
  - Probability of *bl*:

$$\mathrm{P}(bl) = \frac{\mathrm{Count}(ba)}{\mathrm{Count(all\ biphones)}}$$

  - Transitional probability from *b* to *l*:

$$\mathrm{P}(l|b) = \frac{\mathrm{Count}(bl)}{\mathrm{Count(all\ } b\text{-}\mathrm{initial\ biphones)}}$$

- Probability of a word [blɪk]
  - Joint probability (product) of biphones
  - Average probability of biphones
  - Many other possibilities. . .

Biphone probabilities

Biphones are not enough

- Versions of simple biphone probabilities can do reasonably well modeling acceptability of monosyllabic non-words made up of attested sequences (Bailey and Hahn 2001, and others)
- Literal biphones cannot distinguish among unattested sequences
    - $\mathrm{P}(bn) = \mathrm{P}(bd) = 0$

Feature-based generalization

Generalization to novel sequences using phonological features

- Halle (1978, attributed originally to Lise Menn): the *Bach* test
    - Plural of [bax] is [baxs]/*[baxz]/*[baxəz]
    - Generalization according to feature [±voice] of final segment

Feature-based generalization

Generalization to novel sequences using phonological features

- Even without direct evidence about #*bn*, #*bd*, English learners do get evidence about stop+sonorant (or even stop+consonant) sequences, from sequences like *bl*, *br*, *sn*
  - Interpolate: #*br*, #*sn* : $\begin{bmatrix} -\text{syllabic} \\ -\text{sonorant} \end{bmatrix} \begin{bmatrix} -\text{syllabic} \\ +\text{sonorant} \end{bmatrix}$
  - Extrapolate: #*br*, #*bl* : $\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{voice} \\ +\text{labial} \end{bmatrix} \begin{bmatrix} -\text{syllabic} \\ +\text{sonorant} \end{bmatrix}$

Feature-based generalization

Goal of this model

- Learn constraints on possible two-segment sequences, stated in terms of natural classes, and evaluate the amount of support they get from the data

Generalizing from segments to natural classes

Minimal Generalization approach (Albright and Hayes 2002, 2003)

$$
\begin{array}{cccc}
 & \text{b} & \text{l} & \text{u} \\
+ & \text{b} & \text{r} & \text{u} \\
\hline
\rightarrow & \text{b} & \begin{bmatrix} +\text{consonantal} \\ +\text{sonorant} \\ -\text{nasal} \end{bmatrix} & \text{u} \\
+ & \text{g} & \text{r} & \text{u} \\
\hline
\rightarrow & \begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{voice} \end{bmatrix} & \begin{bmatrix} +\text{consonantal} \\ +\text{sonorant} \\ -\text{nasal} \end{bmatrix} & \text{u}
\end{array}
$$

- Input data forms compared pair-wise, extracting what they have in common
- Generalize: Shared feature values are retained, unshared values are eliminated

What to compare with what?

- Comparison between [bl] and [gr] is sort of obvious (they have a lot in common)
- Unfortunately, not all comparisons are so informative
  - By comparing dissimilar clusters, we support very broad abstractions (almost all feature specifications eliminated)
  - E.g., *b*+*s* or *l*+*p* → almost any C

$$
\begin{array}{llll}
 & \text{b} & \text{l} & \text{a} \\
+ & \text{s} & \text{p} & \text{a} \\
\hline
\rightarrow & \begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix} & \begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix} & \text{a}
\end{array}
$$

A potentially fatal prediction

A dangerous misstep: *bl* + *sp* → CC

- In fact, CC clusters are very well attested in English
- Potentially fatal prediction: *bdack* [bdæk] should be very acceptable, because it contains a well-attested sequence:

$$\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix}$$

The challenge

- Find a way to generalize over natural classes such that initial [bl] and [br] provide moderate support for [bn], even though it is outside the feature space that they define
- Prevent comparisons like [bl], [sp] from generalizing to [bd], even though it is within the space they define

Penalizing sweeping generalizations

- Intuitively, [dw] + [gw] : [bw] isn't too great an inductive leap
- This is, in part, because the resulting abstraction is so specific
    - Just need to specify labiality to get [bw]
    - To get [bd] from $\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix}$, we must fill in many features

Penalizing sweeping generalizations

Put differently: $\begin{bmatrix} -\text{son} \\ -\text{cont} \\ +\text{voi} \end{bmatrix} \begin{bmatrix} -\text{syl} \\ -\text{cons} \\ +\text{labial} \end{bmatrix}$ describes a small set of possible sequences (*bw*, *dw*, *gw*)

- If such sequences are legal, the probability of finding any one of them at random is 0.33
- The set of $\begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{lateral} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ -\text{nasal} \\ -\text{strident} \end{bmatrix}$ sequences is large ($\approx 16 \times 11$, or 176 possibilities); the chance of getting [bd] at random < 0.006
- Although both are somewhat supported, the chances of actually encountering [bw] are much higher than [bd]

Learning strategy

- Find descriptions that make 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 & Tesar 2004; Hayes 2004); also related to Bayesian inference, and Maximum likelihood estimation (MLE)

Trying out the intuition

A simple-minded 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

$$\propto \frac{\text{Number of times } xy \text{ occurs in corpus}}{\text{Total number of two-item sequences}}$$

$$\times \text{ Prob(choosing } a \text{ from } x)$$

$$\times \text{ Prob(choosing } b \text{ from } y)$$

Trying out the intuition

Probability of a particular instantiation a of a natural class x

- Simple: $\dfrac{1}{\text{Size of } x \text{ (i.e., number of members)}}$
- Weighted: Relative frequency of $a \times \dfrac{1}{\text{size of } x}$

☞ I will use unweighted values here

Example: probability of [bw]

$$\text{Probability of [bw] using } \begin{bmatrix} -\text{son} \\ -\text{cont} \\ +\text{voi} \end{bmatrix} \begin{bmatrix} -\text{syl} \\ -\text{cons} \\ +\text{labial} \end{bmatrix}$$

$$= \text{Prob}(\begin{bmatrix} -\text{son} \\ -\text{cont} \\ +\text{voi} \end{bmatrix} \begin{bmatrix} -\text{syl} \\ -\text{cons} \\ +\text{labial} \end{bmatrix})$$

$$\times \text{Prob}([\text{b}]|\text{vcd stops})$$

$$\times \text{Prob}([\text{w}]|\text{labial glides})$$

(*relatively high*)

Example: probability of [bw]

$$\text{Probability of [bw] using} \begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{lat} \end{bmatrix} \begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{strid} \end{bmatrix}$$

$$= \text{Prob}(\begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{lat} \end{bmatrix} \begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{strid} \end{bmatrix})$$

$$\times \text{ Prob([b]}|\text{non-nas/lat C's)}$$

$$\times \text{ Prob([w]}|\text{non-nas/strid C's)}$$

(*very low*)

Parsing strings of segments

Given multiple possible ways to parse a string of segments, find the one with the highest probability (Coleman and Pierrehumbert 1997; Albright and Hayes 2002)

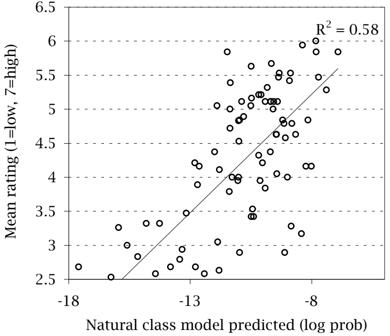
- [bw] can find good support from $\begin{bmatrix} -\text{son} \\ -\text{cont} \\ +\text{voi} \end{bmatrix} \begin{bmatrix} -\text{syl} \\ -\text{cons} \\ +\text{labial} \end{bmatrix}$
- [bd] has no allies that provide such a close fit; it must rely on broader (and weaker) generalizations like $\begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{lat} \end{bmatrix} \begin{bmatrix} +\text{cons} \\ -\text{nas} \\ -\text{strid} \end{bmatrix}$

Local summary

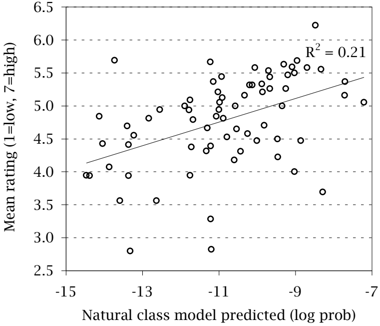
- Procedure for exploring which sequences of natural classes are best supported by the data
- Result: a set of statements about relative likelihood of different sequences

Testing the model

Albright and Hayes (2003) data: ($r(92) = .759$)



Albright (in prep.) data: ($r(70) = .454$)



- Also a reasonable first pass at modeling attested combinations
- Quite a bit better for Albright & Hayes (2003) data
- Fairly even performance across the range of ratings

Summary of this section

- Two different models of gradient acceptability (whole-word similarity, segmental features)
- Both provide decent models of attested sequences
  - See Albright (in prep.) for arguments that feature-based model may be overall better suited to the task
- These results are encouraging, but not too surprising given body of literature showing correlations between acceptability and degree of attestation
- Next: attempt to extend this result to unattested sequences

Preferences among onset clusters

Numerous studies have investigated relative acceptability of unattested clusters using novel words

- Greenberg & Jenkins (1964); Scholes (1966); Pertz & Bever (1975); Coleman & Pierrehumbert (1997); Moreton (2002); Hay, Pierrehumbert & Beckman (2004); Davidson (2006); Haunz (2007); Berent et al., (in press)

Goal of this section

- Examine a selection of findings from this literature, testing the extent to which observed preferences can be predicted by models based on the set of existing clusters

Reason to think some cluster preferences could be learned

Hayes and Wilson (to appear)

- Preliminary demonstration that some preferences among unattested clusters may indeed be learnable
- Trained variety of inductive and similarity-based models on set of existing English onset clusters
- Tested on ability to predict acceptability of novel words, with mix of attested and unattested clusters (Scholes 1966)
- Found impressively good fits ($r > .83$), particularly for their own model ($r = .946$)
  - Plot shown in Bruce's slides yesterday

Strategy here

A first task

- Show that feature-based model also does fairly well on Scholes (1966) data
- Test models on more specific comparisons
  - *#bw* > *#bn* > *#bd* (Berent et al., in press; Albright, in prep)
  - *#bw* > *#dl* (Moreton 2002)
  - *#sp* > *#sk* (Scholes 1966)

Scholes (1966)

- Asked 7th graders about acceptability of words with both attested and unattested onset clusters

plung [plʌŋ], *shpale* [ʃpeɪl], *fkeep* [fkiːp], *ztin* [ztɪn], *zhnet* [ʒnɛt], . . .

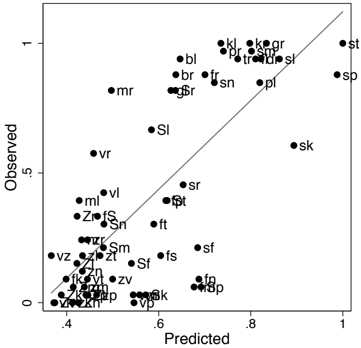
- For each word, counted how many participants deemed possible as an English word—e.g.:

krʌn, stɪn	100%
blʌŋ, slr̩k	84%
glʌŋ, ʃrʌn	72%
nlʌŋ, srʌn	47%
nrʌŋ, zrʌn	33%
vtɪn, fnɛt	19%
ʒpeɪl, vmæt	11%
vkiːp, ʒviːl	0%

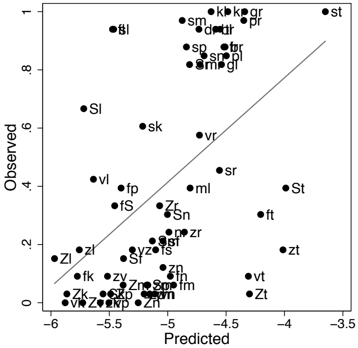
Test 1: whole word acceptability

- Trained models on English words from CELEX
- Tested on Scholes non-words → ratings for entire monosyllable
- Models' predictions rescaled as in Hayes and Wilson

GNM: ($r(60) = .756$)



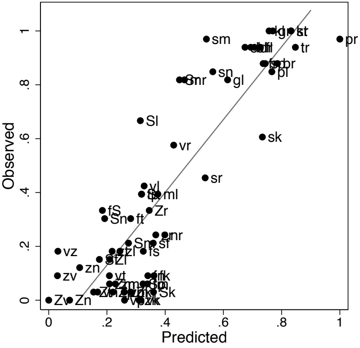
Natural class model: ($r(60) = .503$)



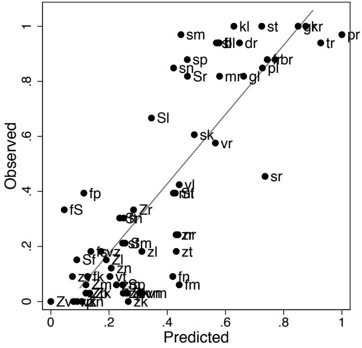
Test 2: Onset acceptability

- Used training corpus from Hayes and Wilson (to appear)
    - Word-initial onsets from CMU pronouncing dictionary, with “exotic” onsets removed (*#sf*, *#zw*, etc.)
- Results are considerably better

GNM: ($r(60) = .881$)



Natural class model: ($r(60) = .830$)



Points to note

- Best results emerge if we assume that subjects based their responses mainly on onset clusters
  - Not implausible! Scholes used just a few rhymes
  - Sendlmeier (1987): subjects focus on salient part of test items
- Even with this assumption, neither model achieves as good a linear fit as Hayes and Wilson's model
  - Bimodal distribution; numerical fits hard to interpret
- Nevertheless, all models make headway in predicting cluster-by-cluster preferences
  - As well or better than on attested sequences
  - If they do this well in predicting other preferences, we'd conclude that there's a good chance they're learnable

Preference for sonority-rising clusters

Berent & al (in press)

- English speakers prefer *#bn* ≻ *#bd*
- As discussed above, this does not appear to mirror any obvious statistical difference between [bn] and [bd] in English

Some new experimental data

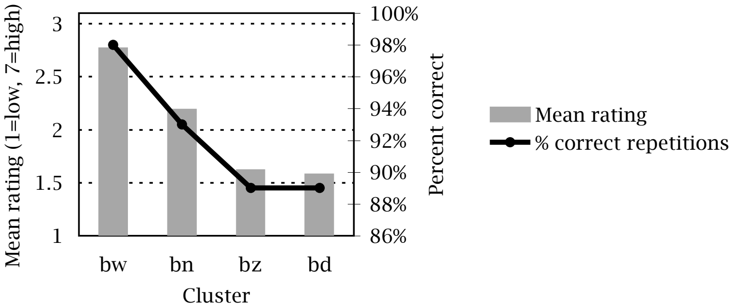
40 non-words with {p, b}-initial clusters

- *#bl*, *#br*, *#bw*, *#bn*, *#bz*, *#bd*; *#pl*, *#pr*, *#pw*, *#pn*, *#pt*
- Paired with variety of rhymes, controlling neighborhood density and bigram probability as much as well (full list in Appendix)
- Rated for plausibility as English words, in same experiment as 70 wug words used above for benchmarking models on attested sequences

Preference for sonority-rising clusters

Results

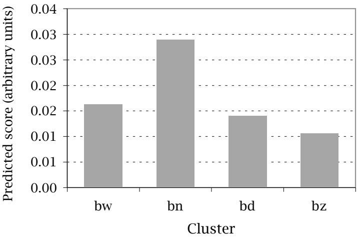
- Acceptability ratings and repetition accuracy both mirror C2 sonority: *#bw* ≻ *#bn* ≻ *#bd*, *#bz*



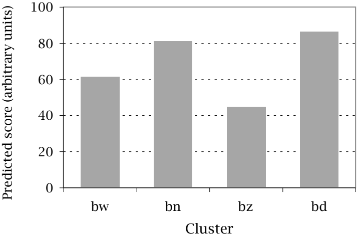
Similarity-based generalization (GNM)

- As above, trained both on whole words (CELEX) and onset corpus (Hayes & Wilson)
- Results: *#bn* ≻ *#bd*, *#bz* predicted correctly over whole words, not onsets

a. Whole word predictions



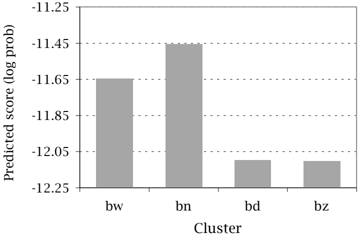
b. Onsets only



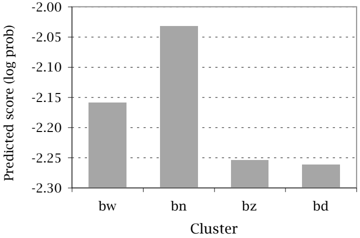
Feature-based generalization (Natural class model)

- In this case, similar predictions regardless of whether trained on whole words or onsets only
- Results are similar as GNM whole-word predictions

a. Whole word predictions



b. Onsets only



Feature-based generalization (Natural class model)

- *#bn* ≻ *#bd*, *#bz* predicted correctly

$$P(\begin{bmatrix} +\text{labial} \\ -\text{son} \\ -\text{contin} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{son} \end{bmatrix}) > P(\begin{bmatrix} +\text{labial} \\ -\text{son} \\ -\text{contin} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{voice} \end{bmatrix}) \quad ☺$$

- *#bw* ≻ *#bn* not correctly predicted

$$P(\begin{bmatrix} +\text{labial} \\ -\text{son} \\ -\text{contin} \end{bmatrix}\begin{bmatrix} +\text{labial} \\ +\text{son} \end{bmatrix}) < P(\begin{bmatrix} +\text{labial} \\ -\text{son} \\ -\text{contin} \end{bmatrix}\begin{bmatrix} +\text{coronal} \\ +\text{son} \end{bmatrix}) \quad ☹$$

Discussion

- Preference for *#bn* ≻ *#bd* tends to emerge from all models
  - Similar preference also predicted by Hayes and Wilson model
  - Berent et al. *bniff* > *bdiff* preference is correctly predicted
- However, #bw ≻ #bn preference not consistently predicted
  - Hayes & Wilson model is the only model to predict direction of preference (Hayes, p.c.)
- Also (not shown): *#pn* ≻ *#pt*, *#ps* not consistently predicted
  - Feature-based bigram model predicts correctly; GNM and Hayes & Wilson models do not

A negative result?

- Models under consideration here capture certain aspects of speaker preferences, but no model consistently predicts full range of preferences
- Must be seen against backdrop of positive results in previous sections
  - All of these models do well on preferences among attested clusters (benchmarking data)
  - Models also make significant headway on unattested clusters, at broad pass (Scholes 1966 data)
  - Failure is specifically in predicting preferences like $\#bw > \#bn$

A negative result?

The failure is interpretable!

- Human ratings reflect preference for stops to be followed by segments that support perceptible bursts, formant transitions (vowels > glides > liquids > nasals > obstruents)

A negative result?

A suggestive result

Although this by no means proves that humans have a substantive bias for stops to be followed by sonorous segments, it shows that current statistical models falter precisely where such biases would be helpful

- The positive payoff of incorporating such a bias will be discussed shortly

A related preference: *#bw* ≻ *#dl*

Moreton (2002)

- Perceptual bias against hearing *#dl* when presented with ambiguous (*dl*∼*gl*) tokens
- Corresponding bias against *#bw* is weaker or non-existent

A related preference: *#bw* ≻ *#dl*

Preference not predicted by any of the models considered here

- Natural class model:

$$P(\begin{bmatrix} +\text{labial} \\ -\text{son} \\ -\text{contin} \end{bmatrix}\begin{bmatrix} +\text{labial} \\ +\text{son} \end{bmatrix}) < P(\begin{bmatrix} -\text{son} \\ -\text{contin} \end{bmatrix}[+\text{approx}]) \qquad ☹$$

- GNM:
  - *#dl* gets support from *#bl*, *#gl*
  - *#bw* gets less support from *#dw*, *#gw* ☹

- Hayes and Wilson model:

$$*[+\text{labial}]\begin{bmatrix} \hat{} & +\text{approx} \\ & +\text{coronal} \end{bmatrix} \gg *\begin{bmatrix} +\text{coronal} \\ -\text{strident} \end{bmatrix}[+\text{cons}] \qquad ☹$$

A related preference: *#bw* ≻ *#dl*

Preference for *#bw* ≻ *#bn* plausibly supported by markedness considerations

- *#dl*/*#gl* more confusable than *#bw*/*#gw* (?)
- Or perhaps an articulatory constraint against coronal+*l* sequences (?)

Upshot of this section

- An initially encouraging result: relative acceptability of *#bn* over *#bd* can indeed be supported by comparison to similar stop+sonorant combinations
- However, this result fails to extend to other combinations with favorable sonority profiles, including *#pn* and *#bw*

A detail from Scholes (1966): *#sp* ≻ *#sk*

#sp: accepted by 29/33 subjects

#sk: accepted by 20/33 subjects

- Unfortunately, different rhymes were used for these two clusters (*spale*, *skeep*)
  - Evidence above suggests rhymes did not influence responses in Scholes' study much
  - If they did, *-ale* should be worse than *-eep* (fewer neighbors, lower bigram probability)

A detail from Scholes (1966): *#sp* ≻ *#sk*

Dispreference for *sk* is found elsewhere, as well

- Cozier (2007): final *-sk* simplified to *-s* in Trinidadian English, but final *-sp* preserved
- Historical change: context-free *sk* > ʃ (OE, OHG)

Testing $\#sp \succ \#sk$

Preference for $\#sp > \#sk$ not predicted by any model tested here

- Nature class-based model: treats both equally

  $$\begin{bmatrix} +\text{coronal} \\ +\text{continuant} \\ -\text{voice} \end{bmatrix} \begin{bmatrix} -\text{continuant} \\ -\text{voice} \end{bmatrix}$$ ☹

- GNM trained on corpus onsets: nearly equal support
  - $\#sp$ (1.0623) vs. $\#sk$ (1.0618) ☹
- Hayes and Wilson model: equal support

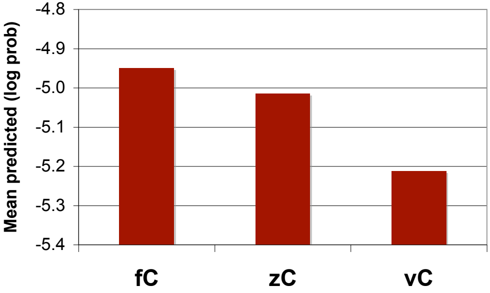
  Both receive perfect scores (no violations) ☹

A possible phonetic basis?

- Cozier (2007): Anticipatory coarticulation from lip closure alters [s] in [sp] → additional cues for stop place
- Greatest benefit in final position, when no following vowel
- Perhaps small added benefit in prevocalic position is nonetheless helpful?

Davidson (2007): #fC ≻ #zC ≻ #vC

- Productions: approx. avg. performance (est. from graphs)
  - #fC: ≈64–70%
  - #zC: ≈57–58%
  - #vC: ≈30–36%
- Trained natural classes on Hayes & Wilson onsets-only corpus
- Tested on *#fn*, *#ft*, *#zn*, *#zt*, *#vn*, *#vt*; averaged over *n*,*t*



Davidson (2007): #fC ≻ #zC ≻ #vC

Points to note

- Predicted order doesn't hold of #C*n* and #C*t* independently
  - #fn ≻ #vn ≻ #zn
  - #zt ≻ #ft ≻ #vt (!!)
- Focusing on /#___ n context, voiceless ≻ voiced predicted successfully, but #zn needs a boost
  - #pl, #pr, #fl, #fr → [labial obstr][coronal sonorant]
  - *#tl, *#θl, *#sr remove support for [anterior obstr][cor son]
  - Plausible boost: advantage of extra amplitude of frication
- For /#___ t context, voicing agreement bias is needed

Despite initially promising results, details may reveal useful role for phonetically motivated biases

Incorporating prior biases

- Results of preceding section have been largely negative (unlearnable preferences)
- In each case, the failure of the models mirrors a possible phoentically-motivated bias
- Goal of this section: demonstrate how incorporating learned statistical generalizations with prior markedness biases can provide a successful overall model
    - Approach follows Wilson (2006): side-by-side comparison of models with/without explicit markedness bias

The sonority bias

- Requirement of interest here: stops should be followed by more sonorous segments
- Plausible restatement in phonetically grounded terms: stops should be followed by segments which...
    - Support perception of burst and VOT
    - Support perception of formant transitions
- These requirements favor following segments which
    - Are strongly voiced
    - Do not interfere with formant transitions, either by blurring/removing them (nasals) or providing independent targets (*l*, *r*, *w*, to varying extents)

The sonority bias

For now, I will treat these as independent requirements

- C_2 sonority: violations reflect availability of voiced formants

	C_2 SON
glide	*
liquid	**
nasal	******
obstruent	*******

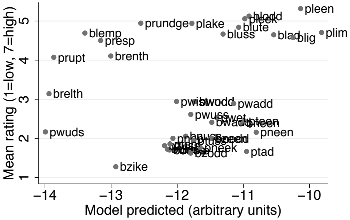
- Non-antagonistic place combinations: violated by *pw*/*bw*, *tl*/*dl*
- Ultimately, may be better combined into a single condition on possible contrasts (Flemming 2004, and others)

Baseline: statistical knowledge or markedness alone

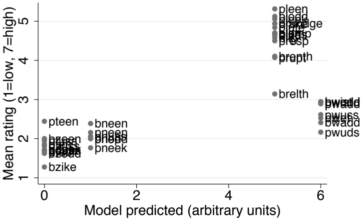
Considered separately, neither the inductive model nor the markedness bias is sufficient to model human preferences

- Statistical model doesn't capture systematic sonority bias
- Markedness bias ignores differences between rhymes

a. Statistical preferences alone ($r(38) = .182$, n.s.)



b. Sonority preference alone

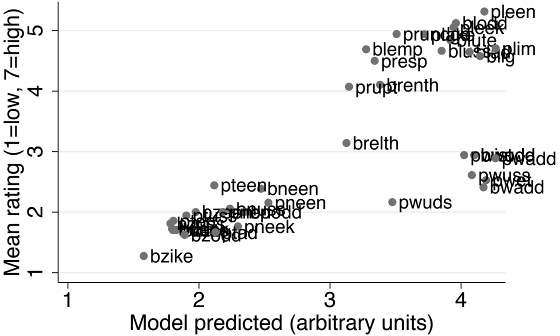


Combining statistical and markedness preferences

- Relative importance of various preferences determined post hoc using a Generalized Linear Model, determining optimal weights (coefficients) by maximum likelihood estimation
- When markedness constraints are added to statistical preferences, a very accurate overall model is obtained

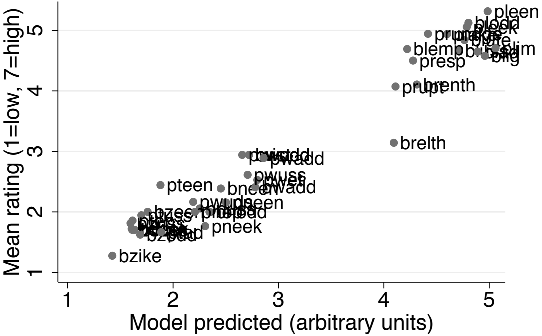
Combining statistical and markedness preferences

Statistical + sonority preferences: $r(38) = .733$



Combining statistical and markedness preferences

Adding $*bw/dl$: $r(38) = .971$



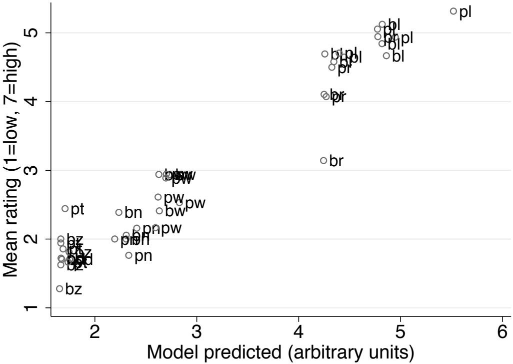
Combining statistical and markedness preferences

Points to note

- Payoff of “sonority jump” from n to l
  - Mimics jump in ratings between #*bl* and #*bn*
  - Possibly just due to attested/unattested difference
  - Happens to correspond to significant difference in availability of formant transitions—perhaps not coincidental that optimal function has this form?
- Bias against lθ], nθ] would further improve fit
  - Items like *prundge*, *brelth*, *brenth* not part of original design
  - Filler items, part of replication of Bailey and Hahn (2001)
  - Included in analysis here for sake of completeness

Combining analogical and markedness preferences

In this case, similar results can be had with combination of analogy + markedness ($r(38) = .969$)



"Experience trumps typology"?

Hammond (yesterday): "Experience trumps typology?"

- Results here show relatively greater effect of phonetic biases, lesser effect of learned statistical generalizations
- Full model:

	Coeff.	Std. Err.	z	Sig.
Stat. model	.2344	.0529	4.43	$p < 0.0001$
C_2 sonority	.5814	.0248	23.47	$p < 0.0001$
OCP	2.4711	.1559	15.85	$p < 0.0001$
Const.	4.4536	.6268	7.11	$p < 0.0001$

The deus ex machina of markedness constraints

- It seems impossible to know a what point a less biased approach is doomed, and when a markedness-based explanation is motivated
- Benefit of attacking the problem from both ends
  - Perhaps revealing that best markedness bias is one that reflects quantitative phonetic differences in availability of cues?
    - Large distance between liquids and nasals
  - Assess concrete performance of combined model, as hand-crafted "standard" for less stipulative models to strive for
- Not a proof that prior markedness bias is required
- Merely a demonstration that current best model is one that incorporates it

The positive result

Two models that do reasonably well on modeling preferences among attested clusters

- GNM and natural class-based model both do fairly well on benchmarking data
- See Albright (in prep) for arguments that natural class model may ultimately be superior

The negative result

Attempts to infer preferences for certain unattested clusters based on attested data: mixed results

- Some preferences evidently inferable given corpus of existing English forms (e.g., $\#bn \succ \#bd$)
- Other preferences are not, at least given currently available models ($\#bw \succ \#bn$, $\#bw \succ \#dl$, $\#sp > \#sk$)

What to conclude?

What do we conclude from this?

- Certainly, it is not possible to exclude the possibility that a better model might succeed where these models have failed
- Many different avenues to explore
  - More refined approaches to evaluating support for combinations of natural classes
  - Different sets of phonological features
  - Different syntax for referring to combinations of segments
  - Not clear whether improvement will ultimately come from incorporating biases directly, as suggested here, or from better feature sets and representations

An unsurprising lesson

Successful statistical models require a good theory of phonology

- Right features and representations
- Right way to apportion "credit" from data to hypotheses (Dresher 2003)
- Right set of prior biases/constraints
  - Externally applied, as in current GLM analysis
  - As part regularization term in constraint weighting (Wilson 2006)

Future directions

Research program outlined here is preliminary attempt to build a framework for comparing and testing hypotheses about these different components

- Broad base of data for benchmarking inductive models with phonological features and representations, but no explicit markedness biases
- Quantitative test of gain from incorporating different pieces of theoretical machinery (Gildea and Jurafsky 1996; Hayes and Wilson, to appear)