Comparative Phonotactics

TWO KINDS OF PHONOTACTICS

1. Absolute phonotactics

- What is the phonological well-formedness of a particular word?
- How is it learned, in the absence of negative evidence?
- Hayes and Wilson (2008) suggested both a grammar framework and a learning system.
  - Framework: the maxent variant (Goldwater and Johnson 2003) of harmonic grammar (Legendre et al. 1990), with the overall constraint-based architecture borrowed from Optimality Theory (Prince and Smolensky 1993)
  - An algorithm selects phonological constraints and weights them so as to
    - maximize the predicted probability of the set of existing words …
    - … against a backdrop of all possible strings.
  - To some extent, this succeeds in matching linguists’ phonotactic descriptions and human phonotactic intuitions.

2. Comparative phonotactics

- Assume two populations of strings, A and B.
- Assume the same maxent framework (constraints, weights, etc.)
- Seek a grammar whose output probabilities accurately predict whether any given string will belong to set A or set B.
- To do this, the constraints must be comparative — make distinctions between the A and B populations.
- Is comparative phonotactics a useful idea for phonology or phonological learnability?

3. Uses of comparative phonotactics — cases I’ll cover

- Analysis of vocabulary strata
  - the Latinate stratum of English
- Discovery of environments for phonological alternations.
  - including: a way to learn (some) opaque phonology
- Discovery of product-oriented generalizations
  - English irregular past tenses
PRACTICAL PRELIMINARIES: HOW I DID THE ANALYSES

4. Data

- All data from English; I used my edited version of the online Carnegie-Mellon Pronouncing Dictionary; transcriptions fixed and all “Level II” forms (Kiparsky 1982) removed.
- The corpus was variously divided into two target populations, as described below.

5. Constraints of the grammar

- I created constraints by hand:
  - from the research literature
  - by scrutinizing comparative populations of all segment unigrams, bigrams, trigrams.
- I used simple search software\(^1\) to assess constraint violations for all words.
- I added and subtracted constraints for my grammars-in-progress, guided by the Akaike Information Criterion.

6. Implementing the maxent grammars

- No need for custom software as in Hayes/Wilson (2008)
- Maxent with just two candidates (population A, population B) is a notational variant of logistic regression, a standard technique of statistical analysis
- I used the `bayesglm()` function of the `arm` package (Gelman et al. 2008, 2009) of the R statistics program (R Development Core Team 2007).

7. Using logistic regression for phonology is not very original!

- Sociolinguists have been using this effective technique for decades, notably with the “Varbrul” program (Cedergren et al. 1974).
- They use it to predict whether an optional rule will apply.
- Here I adopt constraint-based phonology, and seek to show it’s useful for the purposes given in (3).

LEXICAL STRATA

8. The Lexical strata hypothesis

- Chomsky and Halle (1968, 373)\(^2\) proposed that languages with heavy admixtures of loanwords develop synchronically arbitrary lexical strata — groupings of vocabulary that:
  - have a purely diachronic origin (native vs. adapted foreign words)

---

\(^1\) [www.linguistics.ucla.edu/people/hayes/EnglishPhonologySearch](http://www.linguistics.ucla.edu/people/hayes/EnglishPhonologySearch)

\(^2\) A tiny sampling of other work: McCawley (1968), Ito and Mester (1995), Moretone and Amano (1999)
are nevertheless apparent to native speakers as a synchronic phenomenon

- In English the strata are thought to be Native and Learned/Latinate, perhaps with a Greek subdivision of the latter.

9. I think strata are real

- As a native speaker I feel I have a strong sense of the “Latinity” of English words, even though I know no Latin.
- This sense is gradient:
  - very Latinate: protectionism, veterinarian, sexuality, vaporization, industrialization
  - Not Latinate at all: warmth, fresh, swath, shove, pooch, yank, beige, snot
  - Fairly Latinate: palate, oblique, motor, postal, suitor
  - See analysis below, which predicts these distinctions.

10. A research gap?

- To my knowledge phonologists have not attempted to define lexical strata operationally, or establish how they might be learned.³

11. What could constitute the language learner’s evidence for strata?

- Morpheme cooccurrence: if you have -ation, then you likely have con-. (49/613, in my data)
- Morphophonemics: Latinate words undergo different phonological alternation types (SPE)
- Phonotactics: Latinate and native words are phonotactically different.
  - This is just what Ito and Mester (1995) proposed re. the strata of Japanese.

12. Where does the native speaker’s sense of strata come from? A proposal

- They internalize a contrastive phonotactics
  - Population A = native
  - Population B = Latinate
- The contrasting strata are bootstrapped in some way, building up from initially simple information, making use of morphology.
- I’ll cover these aspects in turn — first setting up a grammar from an artificial starting point, then suggesting how it might be bootstrapped.

13. Getting started: an operational definition of Latinity

- Any word of at least seven letters ending in one of these suffixes:⁴

---

³ Just before getting on the plane I noticed Christiansen and Monaghan (2006), which uses what look rather like constraints to distinguish nouns and verb.

⁴ I also required that there be a stem syllable, so that e.g. station did not qualify.
-able, -acy, -al, -ancy, -ant, -ary, -ate, -ated, -ation, -ator, -atory, -ence, -ency, -ent, -graphy, -ia, -iac, -ian, -ible, -ic, -icial, -ician, -ific, -ify, -ine, -ism, -ist, -ity, -ium, -ive, -ize, -ular, -logy, -or, -ory, -ous, -sis, -tion, -ure, -us

- Looking over the data, this seems not too bad to me as an ad hoc way of identifying words that seem Latinate.

14. **Constraints I: those that penalize Latinity**

- Latin had rather stricter phonotactics than English, lacking:
  - Palato-alveolars /ʃ, ʒ, ʒ, ʒ/. These arose later in English by alveolar palatalization, but only in “ambisyllabic” positions (nation, vision, natural, gradual).
  - Initial [sn] (a sound change had turned these all to [n]).
  - No [f] before obstruents ([ft] common in native words).
- Various English sounds just happen not to be the way that Latin sounds normally get rendered; e.g. [u], [au].
- The Latin sounds were transmitted to English in particular ways.
  - [w] is rendered as such only in the clusters [kw] and [gw]; else it appears as [v]; so [w] is missing in other positions.
  - [k, g] undergo Velar Softening to [s, dʒ] before (what used to be) nonlow front vowels ([aɪ, ɪ, ɛ]).
  - Palatal glide [j] is rendered as [dʒ] (except in the diphthong [ju]).
  - *u is [ʌ] before nonfinal coda consonants, else [u] after coronals, else [ju].
- Some English-based phonotactics, like *V before a nonfinal coda consonant, are obeyed with greater strictness in the Latinate vocabulary.
- It’s straightforward to set up constraints based on these factors; e.g. *LATINATE IF [sn

15. **Constraints II: those that penalize nativeness**

- Crudely: just plain length; Latinate words are longer; in our culture we say “long words” when we difficult, rare, learned words.
- Some sound sequences are abundant in Latinate words and not in native words. They sound quite Latinate to me: [VpʃV], [VkʃV], stressless [iə], [mn]
- Even certain individual phonemes are strongly overrepresented in Latinate words: [n], [t]. [v]

16. **The full grammar I set up: Constraints and weights**

<table>
<thead>
<tr>
<th>Prefer Native</th>
<th>Prefer Latinate</th>
</tr>
</thead>
<tbody>
<tr>
<td>INITIAL [sn]</td>
<td>11.84</td>
</tr>
<tr>
<td>MONOSYLLABIC</td>
<td>6.47</td>
</tr>
<tr>
<td>PALATOALVEOLAR Coda</td>
<td>4.08</td>
</tr>
</tbody>
</table>
### Comparative phonotactics

<table>
<thead>
<tr>
<th>Phonological Feature</th>
<th>Weight</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial [ʃ]</td>
<td>3.91</td>
<td>[t]</td>
</tr>
<tr>
<td>Alveolar Stop [l]</td>
<td>2.60</td>
<td>[mn]</td>
</tr>
<tr>
<td>[ɾ̩]</td>
<td>1.77</td>
<td>[ŋ]</td>
</tr>
<tr>
<td>Disyllabic</td>
<td>1.53</td>
<td>At Least 5 Syllables</td>
</tr>
<tr>
<td>w not after [k], [ɡ]</td>
<td>1.39</td>
<td>At Least 4 Syllables</td>
</tr>
<tr>
<td>Precluster Shortening</td>
<td>1.35</td>
<td>[ʃ]</td>
</tr>
<tr>
<td>Final Main Stress</td>
<td>1.24</td>
<td>[əɾə]</td>
</tr>
<tr>
<td>Initial [ʃ] not before [u]</td>
<td>1.23</td>
<td>{[p], [k]}+[ʃ]</td>
</tr>
<tr>
<td>[u]</td>
<td>1.22</td>
<td></td>
</tr>
<tr>
<td>[k,g] + Velar Softening Trigger</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>[ə] in Open Syllable</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Taker of [ju] before [u]</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>General Bias against Latinity (intercept)</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>[ŋ]</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>[θ]</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Trisyllabic Shortening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less Managerial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lengthening</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

#### 17. Computing probability of Latinness for one form: frustration [frasˈtɹæʃən]

- **Frustration** violates four simple constraints penalizing non-Latinity:

  - **Weight**

    | Weight | Probability |
    |--------|-------------|
    | Prefer Latinate if [ŋ] | 0.341 |
    | Prefer Latinate if [t]  | 0.843 |
    | Prefer Latinate if [ʃ]  | 1.610 |
    | Prefer Latinate if Stressless Vowel | 0.119 |

    Total weight 2.904

- **Frustration** violates one constraint penalizing Latinity, the default constraint:

    **General Preference Against Latinity** 0.578

    Total weight 0.578

- The standard maxent formula (e.g. Goldwater and Johnson 2003, (1)) tells us:

  5 [ʃ] per se is actually favored in Latinate words; the preference is overridden by stronger anti-Latinate constraints on [ʃ] that are applicable when it is not in its preferred ambisyllabic position.
• \( P(\text{frustration is Latinate}) = \frac{e^{-0.578}}{1 + e^{-2.904}} = 0.911 \)

• So \textit{frustration} is claimed to be fairly Latinate, but not utterly Latinate.

18. Performance of the Latinity-detecting grammar

• Highest scoring words that I had pre-classified as Latinate (see (13)), all with probabilities at least .996:

  protectionism, veterinarian, sexuality, vaporization, geriatrician, industrialization, perfectionism, reactionary, generalization, pasteurization, popularization, polarization, degenerative, inoperative, insectivorous

• Lowest-scoring words pre-classified as non-Latinate
  ➢ Sampling at random from the bottom 500, all with scores less than .001:

  warmth, fresh, gulch, swath, preach, pooch, yank, beige, snot, munch, scrooge, sniffle, lynch, wont, brooch, width, shrift, should, coach, trench, snub, cringe, drudge, speech

• Lowest-scoring words that I had pre-classified as Latinate
  ➢ This appear almost entirely to be misclassifications, \textit{sardine}.
  ➢ A few are interestingly deviant words with true Latinate suffixes:

  \textit{public} 0.048 [\text{A}] in open syllable
  \textit{wondrous} 0.045 unusual attachment of Latinate suffix to native stem
  \textit{warrior} 0.033
  \textit{vegetable} 0.045 palatoalveolar in coda, due to syncope \[
  \text{[ved3.to.bal]}\]
  \textit{psychic} 0.044 Velar Softening not applied, because Greek (\textit{SPE} suggests a separate sub-stratum for Greek)
  \textit{seismic} 0.034 long \textit{V} in closed syllable, because Greek

• Highest-scoring words preclassified as non-Latinate (all above .975)
  ➢ Most of these seem to be simple misclassifications of my original definition of Latinity (13).
  ➢ A few seem imperfectly Latinate to me and might suggest revisions to the analysis.

  \textit{crucifixion} other spelling of \textit{-tion}
  \textit{Mediterranean, epicurean} other spelling of \textit{-ian}
  \textit{proletariat, secretariat} Suffix I should have included as Latin
  \textit{minutiae} Suffix I should have included as Latin
  \textit{intercession, intermission} Suffix I should have included as Latin
  \textit{verisimilitude} Suffix I should have included as Latin
  \textit{practitioner} stem actually is Latinate
19. Aggregate performance

- For these charts, I separated Latinate and non-Latinate (by my preclassification), then sorted by descending predicted probability.

20. Digression: A theoretical point about the phonotactics of lexical strata

- The Latinity pattern of English is evidence against theories (e.g. Ito and Mester 1995) that assert that the vocabulary strata are nested (native words fill a subset of the phonotactics of the foreign words).
- Here, there is no subset relation in either direction.

---

6 Etymology from OED: “Italian *estravaganza* (an) extravagance (more commonly *stravaganza*), refashioned after Latin *extra*-.”
• The same pattern holds true for Japanese; Kawahara et al. (2005).
• Violations of Ito and Mester’s principle are likely to occur whenever the source and recipient languages are complex in distinct ways.

21. The bootstrapping problem for lexical strata

• No external oracle tells the learner that there is a Latinate stratum at all.
• The distinction must somehow emerge from the language acquisition process.
• But how?

22. A scenario

• For reasons to be made clear, it makes sense for the language learner to collect a contrastive phonotactics like this:
  ➢ Population A = “words that have suffix -x”.
  ➢ Population B = “words that do not have suffix -x”.
• Suppose we (arbitrarily) start doing this with -ation, a common (and canonically Latinate) suffix.
• Look at the forms not ending in -ation that get high scores in the -ation grammar.
  ➢ I checked this and found: these are words that have other Latinate suffixes.
  ➢ This was true for the top 25 words in my list; each ended in a Latinate suffix.
    -ary (6), -ism (5), -al (4), -ist (2), -able (1), -iary (1), -ician (1), -istic (1), -ity (1), -ize (1), -ution (1)
• Repeating the procedure (contrastive phonotactics for -ation plus newcomers) added:
  -ous, -ative, -ator, -ion, -ure, -ent
• Thus we could imaging a bootstrapping operation, gradually uncovering a stratum of affixes that share their contrastive-phonotactic properties.

FINDING THE ENVIRONMENTS FOR PHONOLOGICAL PROCESSES
BY SORTING THE STEM INVENTORY

23. Learning environments by stem-sorting

• Not original with me but proposed by Becker and Gouskova (2012) for Russian data.
• We have some affix that exists in two allomorphic forms a and b.
• We suppose that the stems that take these allomorphs form populations A (“a-takers”) and B (“b-takers”)
• Proposal: language learners perform contrastive phonotactics on the two populations and use the result to distribute the affix allomorphs.

24. Comparison: how this is learned as “pure phonology” in OT

• Adopt some system that learns underlying forms.
• Assume some appropriate set of constraints, perhaps from Universal Grammar (Prince and Smolensky 1993).
• Use a ranking algorithm (e.g. Tesar and Smolensky 2000) that finds the ranking that derives the correct pattern.
• There is no inspection of stems per se.

25. A simple case of stem-sorting

• Hayes, Zuraw, Siptár, and Londe (2009) studied Hungarian vowel harmony.
• Although they didn’t confess this point, our method in practice was precisely that of contrastive phonotactics!
• Two populations of stems:
  ➢ A: those that take front-vowel suffixes
  ➢ B: those that take back-voweled suffixes

26. The most effective way to separate the populations: vowel harmony constraints

• E.g. stems ending in front rounded vowels are always in Population A.
• Those ending in back vowels are always Population B
• Etc.

27. The surprising result

• In the “zones of lexical variation”, where harmony is unpredictable (about 900 stems) stem-final consonants affect harmony.
• More front suffixes when the stem ends in
  ➢ a bilabial consonant
  ➢ a sibilant
  ➢ a coronal sonorant
  ➢ a consonant cluster
• The effect is fairly large: about 1/3 back suffixes when none of these environments is met; close to zero when two are present.

28. Stem-sorting, or ordinary whole-word phonology?

• The vowel constraints work fine as normal phonology — the suffix allomorph that better agree’s with the stem vowel (Lombardi 1999) will surface as the winner.
• But for consonants, things are different — you really have to look at the stems.

29. Why the effect must be a stem effect

• About half of the Hungarian suffixes begin with a consonant in one of the four classes of (27), like dative [-nek]/[-ńok], with a coronal sonorant.
• But these suffixes do not take front allomorphs more often than the others; if anything, it is the reverse.
• Also, the consonant effects on vowel backness fail to show up when you inspect the stem inventory — they are simply not part of Hungarian gradient phonotactics.\footnote{Thanks to Kie Zuraw, who kindly prepared a spreadsheet proving this point when the question arose.}
• To get the distribution right, you must do stem-sorting — just as scenario (23) says.

30. A second application of phonotactic stem-sorting: a common scenario for opaque phonology\footnote{For opacity see Kiparsky (1973); for a review of the (huge) literature see Bakovic (2011).}

• Stem type A takes affix allomorph a.
• Stem type B takes affix allomorph b.
• Then, a phonological process neutralizes the distinction that is used for picking a and b.


• Most consonant stems take 2 sg. /o-/:
  
  \[ \text{[sanga]} \quad \text{‘say-imp.} \quad \text{[o-sanga]} \quad \text{‘say-2 sg.} \]

• Vowel stems take glided [w-]:
  
  \[ \text{[ina]} \quad \text{‘hate-imp.’} \quad \text{[w-ina]} \quad \text{‘hate-2 sg.’} \]

• /b/-stems take /o-/, then b \( \rightarrow \varnothing \ V \quad \_ V \) obscures the output:
  
  \[ \text{[bina]} \quad \text{‘dance-imp.’} \quad \text{/o-bina/} \rightarrow \text{[oina]} \quad \text{‘dance-2 sg.’} \]

• This is standard counterbleeding opacity.
• It is learnable by sorting the isolation stems for whether they take [o-] or [w-].
  \[ \text{[w-]-taking stems always being with a vowel} \]
  \[ \text{[o-]-taking stems always begin with a consonant.} \]

32. Turkish /k/-deletion (Kenstowicz and Kisseberth 1979, 191-193)

• Vowel stems take [-s\(\text{\vold}\)] for 3 sg. poss.:
  
  \[ \text{[arul]} \quad \text{‘bee’} \quad \text{[aru-sul]} \quad \text{‘his bee’} \]

• Consonant stems take [-tu]:
  
  \[ \text{[kuz]} \quad \text{‘daughter’} \quad \text{[ku-zu]} \quad \text{‘his daughter’} \]

• Consonant stems in […]k take [-tu], then lose the /k/ intervocalically:
  
  \[ \text{[ajak]} \quad \text{‘foot’} \quad \text{/ajak-tu/} \rightarrow \text{[ajau]} \quad \text{‘his foot’} \]

• Sorting isolation stems for what allomorph they take solves this problem.

33. Finnish genitive plurals (Anttila 1997)

• Trisyllabic stems ending in /a/ take [-iden]
  \[ /\text{mansikka/} \quad \text{‘strawberry’} \quad \text{[man.si.ko-i.den]} \]

• Trisyllabic stems ending in /o/ take [-jen]
But because of the process $a \rightarrow o / __i$, the difference is not detectible in surface forms.

Allomorphy by stem-sorting could solve the problem.

34. **Prediction**

- If speakers sometimes distribute affix allomorphs using stem-sorting, this particular pattern should be a form of **stable opacity** — unlike contextual counterfeeding in general.

**PRODUCT-ORIENTED GENERALIZATIONS**

35. **Origin of the idea**

- Bybee and Moder (1983)
- Morphological processes can be defined not as an input-output relationship but simply as a phonological characterization of their outputs.
- See Kapatsinsky (2013) for experimental evidence supporting the concept.

36. **Example: English past tenses ending in [ɔt]**


37. **A plausible research agenda**

- Analyze these effects using constraint-based linguistics.
- Expressed product-oriented generalizations as constraints defined on outputs (i.e., specific to a morphological category; not the purely phonological generalizations of OT) and let these constraints participate in the selection of winning candidates.
- See Becker and Gouskova (2012) for application to Russian jers.

38. **What sort of phonotactics should serve as the basis for product-oriented generalizations?**

- I conjecture that **comparative** phonotactics would work better — e.g., Population A = irregular past stems, Population B = all other words
- Why? Consider the past tense of nonce form *pwing*.
  - ??[pwanŋ] has low absolute phonotactic probability, due to its initial cluster.
  - But I judge that it’s a very likely candidate as the past tense of *pwing*.
  - Absolute phonotactics would be fooled here, comparative would not.

39. **Analysis carried out here**

• For simplicity, I used only bare stems; i.e. held, but not beheld.
• I created a simple maxent grammar for comparative phonotactics that distinguishes irregular past stems from ordinary words.

40. The grammar, with examples

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Weight</th>
<th>Forms it prefers</th>
<th>Examples forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE BIAS AGAINST IRREGULARS</td>
<td>10.81</td>
<td>~17000</td>
<td>(almost all words)</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ɛpt]</td>
<td>5.85</td>
<td>6/138</td>
<td>kept</td>
</tr>
<tr>
<td>IRRS. SHOULD BE MONOSYLLABIC</td>
<td>4.37</td>
<td>136/138</td>
<td>most irregulars</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [aʊnd]</td>
<td>4.09</td>
<td>4/138</td>
<td>found</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ɑɪ]</td>
<td>4.02</td>
<td>8/138</td>
<td>taught</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [e] + {{t}, [d]}</td>
<td>3.65</td>
<td>15/138</td>
<td>bet</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ɑŋ]</td>
<td>3.36</td>
<td>8/138</td>
<td>clung</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ɛ] + {{l}, [n]} + {{t}, [d]}</td>
<td>3.28</td>
<td>11/138</td>
<td>felt</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ɪ] + {{t}, [d]}</td>
<td>3.22</td>
<td>12/138</td>
<td>bit</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [æŋk]</td>
<td>2.82</td>
<td>3/138</td>
<td>sank</td>
</tr>
<tr>
<td>IRRS. SHOULD CONTAIN [oʊ]</td>
<td>2.32</td>
<td>22/138</td>
<td>rose</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [u]</td>
<td>2.13</td>
<td>7/138</td>
<td>blew</td>
</tr>
<tr>
<td>IRRS. SHOULD CONTAIN [ʊ]</td>
<td>1.99</td>
<td>5/138</td>
<td>shook</td>
</tr>
<tr>
<td>IRRS. SHOULD CONTAIN [ʌ]</td>
<td>1.96</td>
<td>19/138</td>
<td>wrung, slung</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN [ʊk]</td>
<td>1.68</td>
<td>3/138</td>
<td>took, shook</td>
</tr>
<tr>
<td>IRRS. SHOULD HAVE FINAL STRESS</td>
<td>1.62</td>
<td>138/138</td>
<td>besought</td>
</tr>
<tr>
<td>IRRS. SHOULD END IN ALVEOLAR STOP</td>
<td>0.84</td>
<td>57/138</td>
<td>met, led</td>
</tr>
</tbody>
</table>

41. Some indication that product-oriented generalizations productively govern people’s behavior

• Albright and Hayes’s (2001) nonce-probe experiment: “give us the past tense of the following imaginary verbs.”
• Often, participants would give answers that could not be generated by any rule in the machine-created rule system we had devised (no precedent among existing irregulars for the change the participant made).
• We conjectured that these forms are product-oriented.
• Example: some participants would seize upon a particular product-oriented generalization and stick with it:
  > **Participant 15** (9 of 60 total responses): [baɪz ~ bəʊz], [bredʒ ~ broʊdʒ], [tʃəʊmd ~ tʃəʊʊnd], [dɑɪz ~ dəʊz], [kɑːv ~ kəʊv], [prik ~ prəʊk], [rænt ~ rəʊnt], [swəkɪl ~ skəʊʊl]
Participant 3a (24 out of 60): [blɪg ~ blɑg], [tʃaɪnd ~ tʃʌnd], [dɒrɪs ~ drʌs], [drɪt ~ drʌt], [flet ~ flʌt], [flɒdʒ ~ flʌdʒ(d)], [ɡɛz ~ ɡɑz(d)], [ɡlʌp ~ ɡlʌp(t)], [ɡrɑnt ~ grʌnt], [krv ~ kʌv(d)] etc., etc.

42. Evaluating the grammar in (40) using these data

- I consider only volunteered forms that Albright and Hayes’s rule-based grammar was unable to generate from the present stem — so unlikely to be source-oriented.
- Predictions for the comparative-phonotactic analysis.
  - Real irregular past stems should get relatively high probabilities (the grammar should work for the data it was trained on).
  - Random real words should get low values.
  - The volunteered forms from the Albright/Hayes subjects, if they are following a product-oriented generalization, should get values in the neighborhood of the real irregulars.
43. **Graphs:** actual probabilities
44. **Graphs: log probabilities (reveal differences in the long tails)**

- It looks like the Albright/Hayes subjects were indeed following output-oriented generalizations, and that these can be located in part by performing comparative phonotactic analysis.

**SUMMING UP**

45. **Three possible uses of contrastive phonotactics**

- Lexical strata
- Learning alternations by stem-sorting
- Learning alternations by apprehending product-oriented generalizations
WHAT STILL NEEDS TO BE DONE?

46. **Assess this model against Becker and Gouskova’s approach**

- Becker and Gouskova (2012) believe: comparative phonotactics is deduced from absolute phonotactics.
  - Learn the absolute phonotactics of Population A
  - Learn the absolute phonotactics of Population B
  - Then, probability that a form $x$ belongs to A is

$$\frac{x’s\ phonotactic\ probability\ construed\ as\ A}{x’s\ phonotactic\ probability\ as\ A + x’s\ phonotactic\ probability\ as\ B}$$

- This strikes me as intriguing but oblique — why not solve the problem as directly as possible? The non-contrastive information will probably just be noise.
- My own efforts at applying the BG method to Latinity do yield less accurate results, as measured by summed log probability.

47. **Experimental work with native speakers**

- Consult native intuition on all of these issues (e.g., how Latinate is this word?) with experiments.
- I predict that, e.g. *wepechation* should sound much less Latinate than, say *tenecation*

48. **Better constraint selection**

- What is the right way to find the best constraints?
- This is a hard problem for all constraint-based learnability study.

49. **Bootstrapping**

- Find a mathematically principled and reliable way to bootstrap lexical strata.

References


Becker, Michael and Maria Gouskova. 2012. Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms., SUNY Stony Brook and NYU.


