Some Inductive Methods for Phonological Learning

1. Scenario

- If we had the right resources:
  - the right theory of Universal Grammar (UG)
  - the right theory of learning
  - computational resources and time comparable to what human children have

- Then we could implement a model human, which, given input data comparable to what a human child gets, would acquire morphology and morphophonemics.
  - It would generate new members of paradigms: “if [glɛmp] is a verb, its past tense must be [glɛmpt]”
  - It would produce multiple responses and ambivalent judgments where people do: “past of [splɪŋ] could be [splʌŋ], or [splɪŋd], or perhaps even [splæŋ]”
  - It would learn the system in the same order children do, and make the same mistakes along the way.

- This will happen some day, I think. Should we start now?

2. Albright & Hayes’s Project

- Our learner (Albright and Hayes 2002) inputs a set of base-derived pairs, e.g. for English past tenses:
  
  \[
  \begin{align*}
  \text{[mis]}_{\text{pres}} & \sim \text{[mist]}_{\text{past}} & \text{‘miss(ed)’} \\
  \text{[pres]}_{\text{pres}} & \sim \text{[preʃt]}_{\text{past}} & \text{‘press(ed)’} \\
  \text{[læf]}_{\text{pres}} & \sim \text{[læft]}_{\text{past}} & \text{‘laugh(ed)’} \\
  \text{[hʌɡ]}_{\text{pres}} & \sim \text{[hʌgd]}_{\text{past}} & \text{‘hug(ged)’} \\
  \text{[ræb]}_{\text{pres}} & \sim \text{[ræbd]}_{\text{past}} & \text{‘rub(bed)’} \\
  \end{align*}
  \]

- It learns a grammar that can project derived forms for any given base, with a well-formedness score.

- For past tense of \textit{scride} the grammar learned by the version of the model in Albright and Hayes (submitted) generates three outputs, which match fairly well to native intuitions gathered in a “Wug” test (Berko 1958):

---

1 This talk describes work carried out in collaboration with Adam Albright of UC Santa Cruz. Work was supported by grant BCS 9910686 from the National Science Foundation.
### Table

<table>
<thead>
<tr>
<th>Output</th>
<th>Model’s Prediction (1 = worst, 7 = best)</th>
<th>Mean Rating for 21 subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>scrid</td>
<td>4.12</td>
<td>3.57</td>
</tr>
<tr>
<td>scrided</td>
<td>4.58</td>
<td>4.17</td>
</tr>
<tr>
<td>scrode</td>
<td>4.98</td>
<td>4.39</td>
</tr>
</tbody>
</table>

- Across a variety of Wug words, the model mimics human intuitions reasonably well:
  - Correlation for regular outputs: \( r = .745, \) \( p < .0001 \)
  - Correlation for irregular outputs: \( r = .570, \) \( p < .0001 \)
  
  See Albright (2002) and Albright, Andrade, and Hayes (2001) for further cases.

### 3. Strategies in Learnability Theory

#### I. Rich UG, Simple Learning Theories

Example: Tesar and Smolensky 2000, assuming
- Optimality Theory
- a simple constraint ranking algorithm (Recursive Constraint Demotion)
- a very rich UG, with an innate constraint set sufficient to treat all phonology

#### II. Impoverished UG, Powerful Learning Mechanisms


#### III. Inductive-Baseline Approaches

- Give primacy to inductive methods, with simple UG.
- See where they fail, and then see if the failure is avoided by adding in UG principles—hence, providing support for the existence of these principles.
- Pioneers of this approach: Gildea and Jurafsky 1996
- We aspire to belong to this category

### 4. Two Possible Advantages of Inductivism

- If languages allow arbitrary phonology—i.e. phonology that resist analysis under any reasonable universal constraint set—then we must explain how it is learned (see Blevins 1997, Hayes 1999 for possible cases of arbitrary phonology).

- Inductive study of phonology, especially by machine, uncovers generalizations not previously noticed by linguists, which emerge as psychologically real when tested:
  - All verbs in English that end in voiceless fricatives are regular (Albright and Hayes, submitted)

People are, perhaps, excellent inductive sponges for phonology.
5. Outline of the Talk

- A method for learning “long-distance” environments (trigger nonadjacent to segment)
- Rule generality as a crucial element in inductive learning
  - using stochastic OT and the Gradual Learning Algorithm (Boersma 1997; Boersma and Hayes 2001) as a tool for enforcing generality
- Tentative defense on learnability grounds of a UG principle—the vowel projection

I. LEARNING LONG-DISTANCE ENVIRONMENTS

6. Navajo Sibilant Harmony (Sapir and Hoijer 1967)

- Example: the s-perfective prefix /si-/ is realized as

  ➢ [ši-] if the **first segment** of the stem is a [–anterior] sibilant ((č, ć’, čʰ, š, ž))
  
  /si-čid/ → [ši-čid] ‘he is stooping over’

  ➢ Either [ši-] or [si-] if **somewhere later in the stem** is a [–anterior] sibilant
  
  /si-té:ž/ → [ši-té:ž], [si-té:ž] ‘they two are lying’

  ➢ [si-] otherwise
  
  /[si-tí]/ → [si-tí] ‘he is lying’

7. Representations Assumed By Our System

- A fairly standard feature set.
- Representations and rule environments are sequences of distinctive feature matrices, as in Chomsky and Halle (1968)
- Elements in environments may be specified as optional: ( )
- Elements in environments may be specified as permitted in any number: ([+F])* matches any number of segments in the target form that are [+F].

8. Input to the Learner

- A set of stem ~ perfective pairs:

  [tàš] ~ [šítàš] [gàn], [sigàn] [čʰ’ojin], [šičʰ’ojin]
  
  [tí] ~ [šítí] [si?:?] ~ [sisí?:?] [bà?:?] ~ [sibà?:?]
  
  [č’ìl] ~ [šić’ìl] [kéšgà:], ~ [šıkéšgà:] etc.
  
  [tì:ž], [štì:ž] [k’az] ~ [sìk’az]

- In cases of free variation: one copy of each variant

---

2 See below, (18), for the status of our learning data.
9. Our General Approach to Finding Environments

I. Generalize bottom-up from the lexicon to find candidate environments.
II. Use an evaluation metric to decide which environments to keep.

10. Start: Parse Into Morphemes; Group Forms by Change

<table>
<thead>
<tr>
<th>I. Prefix [sì-]</th>
<th>II. Prefix [šì-]</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. [tí]</td>
<td>a. [tâš]</td>
</tr>
<tr>
<td>b. [bà:?]</td>
<td>b. [t̥eːž]</td>
</tr>
<tr>
<td>c. [gàn]</td>
<td>c. [㎏sg̊ː]</td>
</tr>
<tr>
<td>d. [sìː?]</td>
<td>d. [ɕh̊ːjin]</td>
</tr>
<tr>
<td>e. [k’áz]</td>
<td>e. [ɕ’iɭ]</td>
</tr>
</tbody>
</table>

11. For Each Change, Try to Find an Environment

- Building up incrementally:
  - Step 1: Treat each learning pair as a “word-specific rule,” by separating out the changing part from the invariant part:
    a. $\emptyset \rightarrow \text{sì} / [\text{___ tã \text{ò} š}]$
    b. $\emptyset \rightarrow \text{sì} / [\text{___ t̥eːž}]$
    c. $\emptyset \rightarrow \text{sì} / [\text{___ kéšg̊ː}]$
  - Step 2: Compare pairs of rules which both attach [šì-] (or both attach [sì-]), and extract what their environments have in common, to form a generalized rule.
  - Step 3: Iterate the process, so that ever more general rules get discovered.

12. Comparing Pairs of Rules

Starting with two word-specific rules:

$\emptyset \rightarrow \text{sì} / [\text{___ tã \text{ò} š}]$
$\emptyset \rightarrow \text{sì} / [\text{___ t̥eːž}]$

we collapse them together (details below), using features:

$$
\emptyset \rightarrow \text{sì} / \left[ \begin{array}{c}
\text{___} \\
\text{t} \\
\text{ação} \\
\text{ș}
\end{array} \right]
$$
$$
+\emptyset \rightarrow \text{sì} / \left[ \begin{array}{c}
\text{___} \\
\text{t̥} \\
\text{eː} \\
\text{ž}
\end{array} \right]
$$

$$
=\emptyset \rightarrow \text{sì} / \left[ \begin{array}{c}
\text{___} \\
\text{sonorant} \\
\text{continuant} \\
\text{spread gl.} \\
\text{+anterior}
\end{array} \right]
\begin{bmatrix}
\text{+syllabic} \\
\text{+high} \\
\text{−round} \\
\text{−anterior}
\end{bmatrix}
\begin{bmatrix}
\text{+continuant} \\
\text{−sonorant} \\
\text{−strident}
\end{bmatrix}
$$
• This particular rule looks very unpromising—but with further generalization, the same process arrives quickly at the right answer (below).
• But what should be collapsed with what? [tâš], [têːž] seems obvious, but what of (say) [cʰòːjìn], [cʰ’i?]?
• This is a problem that was much less difficult in our earlier work, which didn’t aspire to learn non-local environments.

Proposed solution:

13. Similarity-Based Alignment

• Here is an intuitively good alignment:

<table>
<thead>
<tr>
<th>čʰ</th>
<th>òː</th>
<th>j</th>
<th>i</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>č’</td>
<td>i</td>
<td>ʃ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Good alignments have two properties:
  ➢ They match phonetically-similar segments.
  ➢ They avoid leaving too many segments unpaired.

• We use existing methods to find the optimal alignment:
  ➢ Theory of segment similarity: Frisch, Broe and Pierrehumbert (1997)
  ➢ Algorithm to search all possibilities for best total similarity (Kruskal 1983)

14. Rationale of Similarity-Based Alignment

• The environments of phonological processes are based on natural classes
• The members of a natural class are phonetically similar
• Thus, aligning similar segment thus favors hypotheses that detect relevant natural classes.
15. Rule Generalization By Collapsing Aligned Pairs of Forms

- Align the forms optimally as described above, and collapse.
- Three rules of generalization:

1. Shared material is collapsed using the feature system.

\[ \emptyset \rightarrow \text{št} / [ ___ k \text{é} \text{š} k\emptyset \] \]

2. Unmatched material is designated as optional, notated with parentheses.

\[ + \emptyset \rightarrow \text{št} / [ ___ t \text{å} \text{š} ] \]

\[ = \emptyset \rightarrow \text{št} / [ ___ [ –sonorant –contin –spread gl. –constr. gl.] [+syllabic –high –round ] \text{š} \] \]

3. Sequential optional elements are collapsed into a single variable, encompassing all of their shared features (e.g. \([+F]^*\)).

\[ \emptyset \rightarrow \text{št} / [ ___ [ –sonorant –contin –spread gl. –constr. gl.] [+syllabic –high –round ] \text{š} \] \]

- When collapsing across two segments that have no features in common, we use the SPE notation \([+\text{seg}]\).
- Iterate by generalizing with the other words in the training data.
- Periodically trim back the hypothesis set, keeping only those rules that perform best.\(^3\)
- Learning terminates when no new “keeper” rules are found.

\(^3\) Specifically: (a) for each word in the training set, keep the most reliable rule (in the sense of Albright and Hayes 2002) that derives it; (b) for each change, keep the rule that derives more forms than any other.
16. Finding the Environment for Nonlocal Sibilant Harmony By Iterative Generalization

\[
\begin{align*}
[\text{č}^h \text{ò: j i n}] & \quad [\text{č'} \text{i ū}] \\
[+\text{sibilant}] & \quad [-\text{continuant}] & \quad [-\text{anterior}] & \quad [+\text{son}] & \quad (-\text{cons}) & \quad (-\text{nasal}) & \quad [-\text{syllabic}] & \quad [-\text{sylabic}] & \quad [-\text{anterior}] \\
[\text{č}^h \text{ǐ t ū}] & \quad \text{[ž ì ū]} & \quad \text{[č}^h \text{ǐ t į]} \\
[+\text{sibilant}] & \quad [-\text{continuant}] & \quad [-\text{anterior}] & \quad ([+\text{seg}])^* \\
[+\text{sibilant}] & \quad (-\text{cons}) & \quad (-\text{nasal}) & \quad [-\text{syllabic}] & \quad [-\text{syllabic}] & \quad [-\text{anterior}] & \quad ([+\text{seg}])^* \\
[+\text{sibilant}] & \quad (+\text{son}) & \quad (+\text{cons}) & \quad (+\text{nasal}) & \quad (+\text{syllabic}) & \quad (+\text{syladic}) & \quad (+\text{anterior}) & \quad ([+\text{seg}])^* \\
\emptyset & \rightarrow & \text{šī-} & / & [\_\_\_\_] & \quad ([+\text{seg}])^* & \quad (\_\_\text{[+\text{seg}]})^* & \quad (\_\_\text{[+\text{seg}]})^* & \quad (\_\_\text{[+\text{seg}]})^* \\
\end{align*}
\]

17. Reexpressing in Standard Notation

- Non-required adjacency to word edge is expressed as null; ([+seg])* as X; hence (16) is:

\[
\emptyset & \rightarrow & \text{šī-} & / & [\_\_\_\_] & \quad ([+\text{seg}])^* & \quad (\_\_\text{[+\text{seg}]})^* & \quad (\_\_\text{[+\text{seg}]})^* & \quad (\_\_\text{[+\text{seg}]})^* \\
\]

TESTING THE APPROACH: A SIMULATION

18. Training Set

- Lacking a morphologically-parsed Navajo data corpus, we are using Pseudo-Navajo…
- 200 whole Navajo words, taken at random from the electronic version of Young, Morgan, and Midgette (1992). ⁴
- Prefixes were attached following the rules of (6):
  > [šī-] if the “stem” began with a nonanterior sibilant (24 stems).

---

⁴ We repeated the learning process on nine other sets of 200 forms, obtaining similar results each time.
Two copies of the stem, one with [ši-], one with [si-], if the stem contained but did not begin with a nonanterior sibilant (34 stems, 2 copies each)

[si-] otherwise (142 stems)

19. The Correct Environments are Learned (among others)

<table>
<thead>
<tr>
<th>a. $∅ \rightarrow [\text{ši-}] / ____ [+\text{sibilant}] [-\text{anterior}]$</th>
<th>Environment for obligatory local harmony.</th>
</tr>
</thead>
<tbody>
<tr>
<td>b. $∅ \rightarrow [\text{ši-}] / ____ X [+\text{sibilant}] [-\text{anterior}]$</td>
<td>Environment for optional harmony, when the stem includes a nonanterior sibilant that is not initial</td>
</tr>
<tr>
<td>c. $∅ \rightarrow [\text{si-}] / ____ X$</td>
<td>Context free environment, takes [si-] by default</td>
</tr>
</tbody>
</table>

+ 84 others, ignored for now but crucial below.

20. Where We Want To Go

- We need a way to put these environments into a grammar that will use them to generate the correct forms.

21. The Rest of the Answer: First Approximation

- Various scholars (Boersma 1998, Russell 1999, Burzio 2002) suggest that we treat morphological mappings as Optimality-theoretic constraints. Thus the rule

$∅ \rightarrow [\text{ši-}] / ____ [+\text{sibilant}] [-\text{anterior}]$ in perfectives

is stated thus as a constraint:

“USE [ši-] / ____ [+\text{sibilant}] [-\text{anterior}] to form the perfective”

- This constraint is violated by forms that begin with a [+sibilant] [-anterior] segment, but use something other than [ši-] to form the perfective. For example:

<table>
<thead>
<tr>
<th>Morphological Base</th>
<th>Candidates that obey USE [ši-] / ____ [+sib] [-ant]</th>
<th>Candidates that violate USE [ši-] / ____ [+sib] [-ant]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[šáp]</td>
<td>[ši-šáp]</td>
<td>*[ši-šáp], *[mú-šáp], etc.</td>
</tr>
<tr>
<td>[táp]</td>
<td>all</td>
<td>none</td>
</tr>
</tbody>
</table>
22. Tableaux

If

• we can rank the three “good” constraints in the right way, namely

\[
\text{USE } [\text{si-}] / _-\text{sib}_-\text{ant} \gg \{ \text{USE } [\text{si-}] / _-\text{ant}, \text{USE } [\text{si-}] / _-\text{ant} \}
\]

crucially tied

• and we figure out some way to keep the other 84 constraints out of the picture, we will get the right outputs:

<table>
<thead>
<tr>
<th>/ si-čid /</th>
<th>USE [si-] / _-[-ant]</th>
<th>USE [si-] / _-X [-ant]</th>
<th>USE [si-] / _-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ši-čid</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>* si-čid</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/ si-té:ž/</th>
<th>USE [si-] / _-[-ant]</th>
<th>USE [si-] / _-X [-ant]</th>
<th>USE [si-] / _-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ši-té:ž</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ši-té:ž</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/si-ti/</th>
<th>USE [si-] / _-[-ant]</th>
<th>USE [si-] / _-X [-ant]</th>
<th>USE [si-] / _-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ši-ti</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* ši-ti</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

II. THE NEED FOR GENERALITY AND HOW TO ENFORCE IT

23. The 84 Other Constraints

• They consist largely of very complicated “generalizations” that happen to hold true of the learning data. Example:

\[
\text{USE } si-/ _-[-del rel] (á) [\text{nasal} \quad \text{HiPitch} \quad \text{-round}]^* \quad \text{([-round])}*]
\]

which works for 24 of the 25 forms to which it applies.

• Such constraints are not just ugly but lead to disaster, as soon as you use the grammar to derive forms that were outside the training data, such as hypothetical /taš/:
This constraint, if ranked high enough, will prevent [šì-tàš] from being generated, or else will keep it from being generated in sufficient numbers.

- So we’d like to be able to get rid of such constraints.
- Failing this, the next best thing is to rank them low enough that they are inactive (Prince and Smolensky 1993).
- Two crucial ingredients will be needed.
  - Gradual Learning Algorithm
  - a way to enforce a preference for general constraints (like those in (22))

### 24. The Gradual Learning Algorithm

- The GLA (Boersma 1997, Boersma and Hayes 2001) can rank constraints in a way that derives free variation, and thus is suitable to the Navajo problem.
- It assigns every constraint a value on a numerical scale (“ranking value”), such that
  - Two constraints C₁, C₂ with the same value are freely ranked—one will dominate the other on any speaking occasion with a probability of 50%.
  - Two constraints separated by 10 units (or so) are essentially categorically ranked.
  - All other probabilities are also possible, given appropriate ranking values
- Our first guess: if you just run the constraints through the GLA for ranking, the junk will settle to the bottom.

### 25. Trouble

- Not true! The junk constraints get ranked high, high enough to make a bad grammar.
- We might have anticipated this:
  - The GLA ranks constraints low when they prefer losing candidates.
  - But, within the learning data, our junk constraints generally prefer only winners.

### 26. The “Tailoring” Problem as a Potentially Fatal Flaw of Inductive Systems

- The danger is that you learn a patchwork of small generalizations that happen to collectively cover the learning data.
- Result: the learned grammar performs splendidly on the corpus you fed it, but fails to generalize to new cases.
- Solution: Do something to favor the broad generalizations over the small ones.
27. **Initial Rankings**

- Boersma 1998 suggests that for morphology, initial rankings might best be based on generality—the more general the constraint, the higher it is ranked before learning takes place.
- We have found that, suitably implemented, Boersma’s idea solves our problem.

28. **A Numerical Characterization of Generality**

\[ \frac{\text{number of forms that a constraint applies to}}{\text{total number of forms exhibiting the change that the constraint requires}} \]

29. **Calculating Generality in the 200-Word Navajo Simulation**

<table>
<thead>
<tr>
<th>Constraint</th>
<th>(a) Relevant forms</th>
<th>(b) Forms with this change</th>
<th>(c) Generality = (a)/(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USE [ši-] / +sibilant +anterior</td>
<td>24</td>
<td>58 [ši-] forms</td>
<td>.414</td>
</tr>
<tr>
<td>USE [ši-] / +sibilant +anterior</td>
<td>58</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>USE [si-] / ___ X</td>
<td>176</td>
<td>176 [si-] forms</td>
<td>1</td>
</tr>
<tr>
<td>Constraint (25) (“junk” constraint)</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

30. **Making Sure Generality Will Make a Difference**

- Rescale so that the original range of generality is converted to a very large range (0-500) on the GLA ranking scale.

For each constraint \( c \), initial ranking value \( v_c = 500 \times \frac{\text{Generality}_c - \text{Generality}_{\text{min}}}{\text{Generality}_{\text{max}} - \text{Generality}_{\text{min}}} \)
31. Rerunning the GLA with Generality-Based Initial Rankings

<table>
<thead>
<tr>
<th>Generality</th>
<th>Initial Ranking</th>
<th>Final Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>550</td>
<td>546.9</td>
<td>550</td>
</tr>
<tr>
<td>1, 1</td>
<td>500, 500</td>
<td>500, 500, 499.98</td>
</tr>
<tr>
<td>.9</td>
<td>450</td>
<td>450</td>
</tr>
<tr>
<td>.8</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>.7</td>
<td>350</td>
<td>350</td>
</tr>
<tr>
<td>.6</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>.5</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>.4</td>
<td>190.1</td>
<td>200</td>
</tr>
<tr>
<td>.3</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>.1</td>
<td>47.6</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

32. Final Ranking Obtained (Hasse Diagram)

- All other junk constraints are likewise well below the top three.
- As far as active constraints go, this is just what we wanted: the ranking of (22).
33. Why Did the Generality-Based Initial Ranking Work?

- Junk constraints start low, and they stay there. Reason:
  - The GLA is error-driven.
  - The general constraint that does the same work as the junk constraint just needs a little time to get ranked correctly, at which point the errors that would promote the junk constraint cease to be made.
- Good constraints with specific contexts, like “USE [ši-] / [+sib]”, are also nongeneral—but appropriately so.
  - They start low, but they are needed to avert errors like *[ši-šáp], so they are soon promoted by the GLA to the very top of the grammar.

III. THE INDUCTIVE BASELINE: VOWEL HARMONY

34. Vowel Harmony

- Vowel harmony is a very characteristic nonlocal process: vowels “see” each other across strings of intervening consonants.
- Not all harmony systems have “consonant invisibility” (cf. Turkish, Clements and Sezer 1982), but most do.

35. Theoretical Proposals

- Vergnaud and Halle (1979): “vowel projection”

  “A projection is a representation, simultaneous with the ordinary phonological representation, on which only those elements are present that share some well-defined phonetic or structural characteristic.” (McCarthy 1979, 30-31)

  \[ \begin{array}{c}
  \text{ʃ} \\
  \text{o} \\
  \text{ʃ}\
  \end{array} \] \begin{array}{c}
  \text{ʃ} \\
  \text{o} \\
  \text{ʃ}\
  \end{array} \] \begin{array}{c}
  \text{o} \\
  \text{o} \\
  \text{o}\
  \end{array} \]  Hungarian “chauffeur”

  \[ \begin{array}{c}
  \text{j} \\
  \text{o} \\
  \text{ʃ}\
  \end{array} \] \begin{array}{c}
  \text{o} \\
  \text{o} \\
  \text{o}\
  \end{array} \]  vowel projection of [ʃoʃɔr]

- Archangeli and Pulleyblank (1987): maximal and minimal scansion

- If the language learner has access to a vowel projection, do otherwise-inaccessible harmony systems become accessible?
36. **Vowel Harmony in Hungarian**

- see Siptar and Törkenczy (2000) for a good overview of the data and literature
- vowels of Hungarian:

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Neutral</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>y, y:</td>
<td>i, i:</td>
<td>u, u:</td>
</tr>
<tr>
<td>Mid</td>
<td>ø, ø:</td>
<td>e:</td>
<td>o, o:</td>
</tr>
<tr>
<td>Low</td>
<td>ε</td>
<td>ø, a:</td>
<td></td>
</tr>
</tbody>
</table>

I will call these F, N, and B; thus [pɔpiːr] ‘paper is a “BN word”

37. **Traditional Story**

- If a single stem vowel is Front or Back, it governs harmony:
  
  - [tyːz-nɛk] F ‘fire-dative’
  - [haːz-nɔk] B ‘house-dative’

- In cases of conflict, the closest vowel determines the outcome:
  
  - [ʃoːʃɛr-nɛk] BF ‘chauffeur-dative’
  - [ɔmɔːbɔ-ɔnɔk] BFB ‘amoeba-dative’

- If a stem consists of all Neutral vowels, it takes a front suffix.  
  
  - [seːʃɛn-nɛk] NN ‘poor-dative’

- A stem ending in a Back plus a sequence of Neutrals takes back harmony:
  
  - [pɔpiːr-nɔk] BN ‘paper-dative’
  - [haʃiːtɛk-nɔk] BNN ‘slot-dative’

38. **A More Articulated Story**

(Siptár and Törkenczy 2000, 70-74)

- Hungarian neutral vowels are not transparent, but **translucent**, triggering back harmony (either obligatorily or in free variation) in a number of words.
- Height effect: the lowest neutral vowel, [ɛ], is opaque far more often than the others.

  - [ɔkteroːr-nɛk] BBN ‘October-dative’

---

5 With the surprising exception of the exceptional class of “hid” words, not covered here.
39. Empirical Survey

- All 708 nouns ending in BN from an electronic dictionary\(^6\) were intuited for their harmonic behavior by Zsuzsa Londe, a fluent native speaker of Standard Hungarian.

<table>
<thead>
<tr>
<th>Vowel Sequence</th>
<th>Back</th>
<th>Vacillates</th>
<th>Front</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. …Bi</td>
<td>315</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>bi</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>c. …Be:</td>
<td>126</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>be:</td>
<td>3</td>
<td>189</td>
<td>7</td>
</tr>
</tbody>
</table>

40. Goal

- Construct a grammar that stochastically predicts the propensity towards back/front harmony for each stem type.

41. Pseudo-Hungarian

- Rather than plunging in to the whole language, we’ve been working up through simulations of increasing complexity, using versions of Pseudo-Hungarian—I will describe three such simulations.

- Vowels of Pseudo-Hungarian:

<table>
<thead>
<tr>
<th>Front</th>
<th>Neutral</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>y</td>
<td>i</td>
</tr>
<tr>
<td>Mid</td>
<td>e</td>
<td>o</td>
</tr>
<tr>
<td>Low</td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>

- Consonants of Pseudo-Hungarian = English: initial, medial, and final consonants and consonants clusters occur in each learned set in proportions representative of their frequency in English.

42. At Issue

- Can the influence of vowel height on harmony be learned even though the vowels are separated by consonants?

43. Simulation 1

- Vowels: all six vowels as monosyllables; all 36 possible combination as disyllables.

\(^6\) Authors unknown; available from http://sourceforge.net/projects/wordlist-hu/.
• Consonants:
  ➢ Fill in the slots: [ __ V __ ] and [ __ V __ V __ ] with initial, medial, and final clusters of English,
  ➢ Frequencies matched the English frequencies of these clusters.
  ➢ Each vowel sequence was filled in with 10 consonant sequences. Hence \((6 + 36) \times 10 = 420\) words in the learning set.

• Vowel harmony: one suffix [-a/-e], which is
  ➢ back if the rightmost stem vowel is back
    \([\text{daflar}] \sim [\text{daflar-a}]\)
  ➢ front if the rightmost stem vowel is \([y]\)
    \([\text{fufyns}] \sim [\text{fufyns-e}]\)
  ➢ front if all vowels are front
    \([\text{witet}] \sim [\text{witet-e}]\)
  ➢ back if the first vowel is back and the second vowel is neutral
    \([\text{dorek}] \sim [\text{dorek-a}]\)

• Learning procedure: as with Pseudo-Navajo.
  ➢ Find the nonlocal environments
  ➢ Assign initial ranking values based on generality
  ➢ Run the Gradual Learning Algorithm
  ➢ Report only the active constraints

44. Grammar Learned for Simulation 1

516.667 \( \text{USE -e / y C}_0 ___ \) dominates

506.000 \( \text{USE -a / } \begin{bmatrix} V \\ +\text{back} \end{bmatrix} X ___ \) dominates

494.000 \( \text{USE -e / } \begin{bmatrix} V \\ -\text{back} \end{bmatrix} C_0 ___ \)

• This grammar derives all and only correct forms
• The system “learned C\(_0\)”; i.e. it generalized from the actual final clusters in the learning data to the correct conclusion: “ignore consonants”.

45. Simulation 2

• Goal: learn a system in which
  ➢ \([\text{Back + i}]\) words take back harmony 75% of the time
  ➢ \([\text{Back + e}]\) words take back harmony 50% of the time.
• Setup: Same as Simulation 1, except that extra copies of the \([\text{Back + i}]\) and \([\text{Back + e}]\) words are added in, to create the free variation.
46. Result for Simulation 2

- Crash and burn!
- Simple constraints distinguishing [Back + i] and [Back + e] were not learned. Here are the top 8 constraints:

   589.098  USE -e / y C0 ___
   578.473  USE -e / [+cons][–syl][–cons] V([+cons])C(y)[+ant][–low] * ___
   577.458  USE -a / [+back] X ___
   577.204  USE -e / [+back] C0 ___
   575.997  USE -a / (s)[+cont][–lab][–syl] ([–cont])[+syl][–cons] [+cons] * ___
   575.768  USE -e / [C0 [+high]][+cons] V [+back] [+cons] (–voice) * ___
   574.967  USE -e / [(y)[–syl][o] [+back] C0 ___
   573.993  USE -a / C [+back] C(i) ___

- System doesn’t even mimic its training data all that accurately; different consonant clusters result in substantially different degrees of backness.
- Conjectured diagnosis: The consonant clusters imposed too many false alignments, making proper alignments of high and mid vowels too rare to generalize from.

47. Simulation 3

- Assumption: UG makes accessible a representation of the vowel strings of words (Vergnaud and Halle 1979 et al.)
- Implementation: learning data same as Simulation 2, with all the consonants deleted

48. Simulation 3 Result

- There are only three active constraints, all sensible, and ranked very close:

   500.601  USE -a / [+back] (i) ___
   500.067  USE -e / [+back] ___
   499.933  USE -a / [+back] ([–round]) ___
When this grammar is tested (run each input thousands of times), we find that target frequencies are well matched.

<table>
<thead>
<tr>
<th>Stem Type</th>
<th>Stem</th>
<th>Output</th>
<th>Proportion that should be generated</th>
<th>Proportion that was generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>[B + e]</td>
<td>ae,</td>
<td>ae-a,</td>
<td>0.500</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td>oa,</td>
<td>oe-a,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ue</td>
<td>ue-a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[B + i ]</td>
<td>ai,</td>
<td>ai-a,</td>
<td>0.750</td>
<td>0.766</td>
</tr>
<tr>
<td></td>
<td>oi,</td>
<td>oi-a,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ui</td>
<td>ui-a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

plus correct results for all non-free-variation forms

49. Tableaux for Simulation 3

<table>
<thead>
<tr>
<th>/aa/</th>
<th>USE -a / [V +back] (i)</th>
<th>USE -e / [V -back]</th>
<th>USE -a / [V +back] (V -round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*aaa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*aae</td>
<td>*(!)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/ay/</th>
<th>USE -a / [V +back] (i)</th>
<th>USE -e / [V -back]</th>
<th>USE -a / [V +back] (V -round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*aye</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*aya</td>
<td>*!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/ai/</th>
<th>USE -a / [V +back] (i)</th>
<th>USE -e / [V -back]</th>
<th>USE -a / [V +back] (V -round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*aia</td>
<td></td>
<td></td>
<td>*(!</td>
</tr>
<tr>
<td>*aie</td>
<td>*(!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/ae/</th>
<th>USE -a / [V +back] (i)</th>
<th>USE -e / [V -back]</th>
<th>USE -a / [V +back] (V -round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*aee</td>
<td></td>
<td></td>
<td>*(!</td>
</tr>
<tr>
<td>*aea</td>
<td>*(!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

50. Upshot

- for one particular pseudo-language, and
- for one particular inductive learning algorithm
- a UG principle, vowel projection, makes learning possible where it would not otherwise be.
51. Caveats

- Would a better inductive learner have succeed on Similation 2?

52. Summary

- Two mechanisms proposed here:
  - Generalization using similarity-based alignment
  - Initial rankings for the GLA based on generality

appear to have potential as the basis of an inductive-baseline learner

- Long-term goal:
  - Amplify this inductive base, with principles of UG as well as further refinements of learning, to make it approach the state of being an adequate model of phonological learning

References


Albright, Adam and Bruce Hayes (submitted) “Rules vs. analogy in English past tenses: a computational/experimental study,” submitted to *Cognition*. http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/


bhayes@humnet.ucla.edu