Learning hidden structure in paradigms*

- Speakers have a rich and detailed knowledge of their lexicon, which they evidence in their treatment of novel words (“wug-testing”). I will show that this knowledge is biased by naturalness: The same kinds of relations that cause regular processes in some languages, regulate irregular processes in other languages. This means that this lexical knowledge is mediated by the grammar.
- I propose an OT-based model in which regular and irregular morpho-phonology is derived from the same set of universal constraints, \( \mathcal{C} \).

1 The naturalness of lexical trends

1.1 Turkish (Becker, Ketrez & Nevins 2008)


We searched TELL (Inkelas et al. 2000), and found that final stops in mono-syllables mostly don’t alternate, but in poly-syllables they mostly do.

<table>
<thead>
<tr>
<th>Size</th>
<th>( n )</th>
<th>% alternating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosyllabic, simplex coda</td>
<td>137</td>
<td>12%</td>
</tr>
<tr>
<td>Monosyllabic, complex coda</td>
<td>164</td>
<td>26%</td>
</tr>
<tr>
<td>Polysyllabic</td>
<td>2701</td>
<td>59%</td>
</tr>
</tbody>
</table>

Most final \( t \)’s don’t alternate, other stops mostly do.

<table>
<thead>
<tr>
<th>Place</th>
<th>( n )</th>
<th>% alternating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial (p)</td>
<td>294</td>
<td>84%</td>
</tr>
<tr>
<td>Coronal (t)</td>
<td>1255</td>
<td>17%</td>
</tr>
<tr>
<td>Palatal (ʧ)</td>
<td>191</td>
<td>61%</td>
</tr>
<tr>
<td>Dorsal (k)</td>
<td>1262</td>
<td>85%</td>
</tr>
</tbody>
</table>

Two other factors that partially predict alternation: The height and backness of the final vowel of the stem.

<table>
<thead>
<tr>
<th>Height of stem’s final vowel</th>
<th>( n )</th>
<th>% alternating</th>
</tr>
</thead>
<tbody>
<tr>
<td>−high</td>
<td>1690</td>
<td>42%</td>
</tr>
<tr>
<td>+high</td>
<td>1312</td>
<td>72%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backness of stem’s final vowel</th>
<th>( n )</th>
<th>% alternating</th>
</tr>
</thead>
<tbody>
<tr>
<td>−back</td>
<td>1495</td>
<td>50%</td>
</tr>
<tr>
<td>+back</td>
<td>1507</td>
<td>60%</td>
</tr>
</tbody>
</table>

---

*Ideas presented today owe much to discussions with Adam Albright, Wendell Kimper, John McCarthy, Joe Pater, and Matt Wolf. Thanks also to the audience at MUMM 2, especially Edward Flemming, John Kingston, and Donca Steriade, and the audience at the UCSC linguistics department, especially Junko Ito, Grant McGuire, Armin Mester, and Matt Tucker. I assume the responsibility for any remaining errors, here and elsewhere.
We gave 24 Turkish speakers a novel noun task (“wug-test”, Berko 1958) with 72 novel nouns of four places (p, t, ʧ, k), three sizes (CVC, CVCC, CVCVC), and eight vowels (a, i, e, i, o, u, ø, y).

The speakers replicated the size and place effects from the lexicon, as in (6), but not the vowel quality effects (not shown, see stats and more detail in paper).

(6)  

\[ \begin{array}{c|c|c}  
\text{lexicon} & \text{human responses} \\
\hline  
CVCp & 0.0 \\
CVCt & 0.2 \\
CVCk & 0.4 \\
CVp & 0.6 \\
CVt & 0.8 \\
CVk & 1.0 \\
CVCVp & 0.3 \\
CVCVt & 0.5 \\
CVCVk & 0.7 \\
CVCVk & 0.9 \\
\end{array} \]

(7) It’s natural to treat alternations in mono-syllabic stems separately from poly-syllabic stems via initial syllable faithfulness (Beckman 1997, 1998, Casali 1998).

(8) It’s natural to treat the propensity of different stops to voice differently.

(9) However, no language is known to change the voicing of a consonant based on the height or backness of a neighboring vowel.

In other words, Turkish speakers only learned the natural (=typologically supported) aspects of their lexicon, and ignored the unnatural ones. UG acts as a filter on the kinds of generalizations that speakers learn.²

1.3 Dutch (Ernestus & Baayen 2003)

Essentially, the same as Turkish:

<table>
<thead>
<tr>
<th>Imperative</th>
<th>Past tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>stop-tə</td>
</tr>
<tr>
<td>top</td>
<td>top-də</td>
</tr>
</tbody>
</table>

Speakers replicated the lexical trends for the different final obstruents (p, t, s, f, χ). They also replicated the vowel length effect: They preferred voicing alternations for stops that followed long vowels.

In the Dutch lexicon, there are more alternations after high vowels (which are all short) than after non-high short vowels — but speakers did not replicate this trend.

Again, this is natural: Vowel length is correlated with the voicing of a following obstruent in many languages (e.g. English), but vowel height is not.

1.4 A note on methodology

Possible objection: Your dictionary does not represent the knowledge of the people you tested, because it is old/riddled with errors/does not show morphological composition/comes from a different dialect/etc.

Responses:

(11) The trends that speakers project from their data correlate with the lexicons we have remarkably well. Whatever the shortcomings of our dictionaries are, they are still a very good approximation of the real data in Turkish, Dutch, Hebrew, Tagalog, Hungarian, Spanish, etc.

(12) We test the trends in the data with sampling (Baayen 2008), showing that these trends are strong even in lexicons that only share about 63% of their items.

(13) It would be ideal to test each speaker twice: Once on their real lexical items, and once with novel items. This way, there is no one idealized lexicon, but rather an actual, separate lexicon for each speaker. I am working on this with Adam Albright and Andrew Nevins.

²In Hayes et al. (to appear), unnatural trends in the data are learned, but they are attenuated relative to natural trends.
2 Analysis

2.1 Grammar-based analysis

Work "inside out" (Hayes 1995, 1998, 1999), so the alternations are considered to be irregular intervocalic voicing.

(14) The UR’s of [sop] and [ʤop] are /sop/ and /ʤop/
(15) The UR of the possessive is /u/ (actually just a high vowel)
(16) /sop + u/ → [sopu] requires IDENT(voice) ≫ *VpV
    /ʤop + u/ → [ʤobu] requires *VpV ≫ IDENT(voice)


(17) IDENT(voice)\textsubscript{sop} ≫ *VpV ≫ IDENT(voice)\textsubscript{ʤop}

From this point on, every word that is sensitive to the ranking of IDENT(voice) relative to *VpV will be listed:

(18) /top + u/ | IDENT(voice) | *VpV
    a. sop top-u |   | *
    b. sobu     |   | !

(19) /ot + u/ | IDENT(voice) | *VpV
    a. sop ot-u |   | *
    b. od-u     |   | !

(20) IDENT(voice)\textsubscript{[sop, top, alp, ...]} ≫ *VpV ≫ IDENT(voice)\textsubscript{[ʤop, harp, ...]}

Until the speaker gets:

(21) IDENT(voice)\textsubscript{[22 items]} ≫ *VpV ≫ IDENT(voice)\textsubscript{[8 items]}

Novel p-final mono-syllables will have a 8/30 (=27%) chance of alternating with [b].

The result: the lexical statistics are built into the grammar. In other words, the distinction between grammar and lexicon is blurred, so that partially-predictable information is not buried in the lexicon.

2.2 What’s wrong with a UR-based analysis?


(22) The UR’s of [sop] and [ʤop] are /sop/ and /ʤoB/
(23) The UR of the possessive is /u/ (actually just a high vowel)
(24) /sop + u/ → [sopu] requires IDENT(voice) ≫ *VpV

\begin{tabular}{|c|c|c|}
\hline
    sop + u     & IDENT(voice) & *VpV \\
\hline
  a. sopu      &   & * \\
  b. sobu     &   & !
\hline
\end{tabular}

(25) /ʤoB + u/ → [ʤobu] is consistent with IDENT(voice) ≫ *VpV

\begin{tabular}{|c|c|c|}
\hline
    ʤoB + u     & IDENT(voice) & *VpV \\
\hline
  a. ʤopa     &   & (*) \\
  b. ʤobu     &   & (*)
\hline
\end{tabular}

The grammar is consistent: IDENT(voice) ≫ *VpV

The problem: The learner has no way to encode the relative numbers of /p/’s and /B/’s in the grammar. Going directly to the lexicon to find them there, unhindered by UG, will find the vowel quality generalizations that speakers don’t have.

Conclusion: Assume the bases as UR’s, assume that affixes only have segments in them, and try to get everything else by ranking constraints. Clone constraints as necessary.
3 Fallback: When the grammar is not enough

Korean (Albright 2008b):

(26) Unmarked Accusative

<table>
<thead>
<tr>
<th>Word</th>
<th>Form</th>
<th>Meaning</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>nat̚</td>
<td>nasɨl</td>
<td>‘sickle’</td>
<td>375</td>
</tr>
<tr>
<td>nat̚</td>
<td>naʧʰɨl</td>
<td>‘face’</td>
<td>160</td>
</tr>
<tr>
<td>nat̚</td>
<td>natʰɨl</td>
<td>‘piece’</td>
<td>113</td>
</tr>
<tr>
<td>nat̚</td>
<td>naʤɨl</td>
<td>‘daytime’</td>
<td>17</td>
</tr>
<tr>
<td>nat̚</td>
<td>nadɨl</td>
<td>‘grain’</td>
<td>1</td>
</tr>
</tbody>
</table>

Assuming /nat̚/ for the roots and /ɨl/ for the accusative can do some work:

(27) /nat̚+ɨl/ → */VtV I(voice) I(asp)*

| a. | natɨl | *!
| b. | nadɨl | *!
| c. | natʰɨl | *

(28) /nat̚+ɨl/ → [natʰɨl], [naʧʰɨl]
 requires */VtV I(voice) I(asp)*

(29) /nat̚+ɨl/ → [nadɨl], [naʤɨl]
 requires */VtV I(voice) I(asp)*

(30) Ident(voice){113+160+1+17 items} ≫ Ident(asp) ≫ Ident(voice){375 items}

The prediction for a novel form, [paf ]:

(31) 94% chance of [tʰ], [ʧʰ], 6% chance of [d], [ʤ]

But are there plausible constraints that will map /nat̚+ɨl/ to [naʤɨl] or [naʧʰɨl]? It seems awfully hard to palatalize without a front vowel around.

With [natʰɨl] as the intended winner, [natʰɨl] is most faithful to it, but still incurs an Ident(ant) violation → add the missing feature as floating in the UR of the accusative affix: /−ant il/.

(32) /nat̚+il/ → [natʰɨl], [nadɨl]
 requires Max(float) ≫ Ident(ant)

(33) /nat̚+il/ → [natʰɨl], [naʧʰɨl], [nadɨl], [naʤɨl]
 requires Ident(ant) ≫ Max(float)

(34) Ident(ant){113+160+1+17 items} ≫ *TI ≫ Ident(ant){375 items}

The prediction for a novel form, [paf ]:

(35) 56% chance of [s], 44% chance of [tʰ], [ʧʰ], [d], [ʤ]

Summary of the predictions:

<table>
<thead>
<tr>
<th></th>
<th>Ident(cont) vs. *TI</th>
<th>Ident(voice) vs. Ident(asp)</th>
<th>Ident(ant) vs. Max(float)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>56%</td>
<td>61%</td>
<td>25%</td>
</tr>
<tr>
<td>[ʧʰ]</td>
<td>94%</td>
<td>61%</td>
<td>25%</td>
</tr>
<tr>
<td>[tʰ]</td>
<td>44%</td>
<td>39%</td>
<td>16%</td>
</tr>
<tr>
<td>[d]</td>
<td>6%</td>
<td>61%</td>
<td>2%</td>
</tr>
<tr>
<td>[ʤ]</td>
<td>39%</td>
<td>39%</td>
<td>1%</td>
</tr>
</tbody>
</table>

*TI, which wants assimilation before a high vowel (Kim 2001), takes care of [s]:

(36) /nat̚+il/ → [nasɨl]
 requires *TI ≫ Ident(ant)

(37) /nat̚+il/ → [natʰɨl], [nadɨl]
 requires Ident(ant) ≫ Max(float)

(38) /nat̚+il/ → [nasɨl]
 requires *ʃ ≫ Ident(ant), Max(float)

(39) *ʃ ≫ Ident(ant){113+160+1+17 items} ≫ Max(float) ≫ Ident(ant){375 items}

(40) 61% chance of [ʧʰ], [ʤ], 39% chance of [tʰ], [d]
The high probability of \[s\] and \[ʧʰ\] conforms with the report in Albright (2008b) about the treatment of novel forms, loanwords, and many native items.

My analysis expresses the language-specific frequencies of mappings in terms of rankings of universal constraints.

4 Last resort: Suppletion and diacritics

It’s certainly not the case that every paradigmatic relation can be derived with phonological mechanisms, e.g. English go \(~\) went.

English \(ɔt\)-takers: teach, catch, think, bring, seek, fight, buy — how many of those can map to their past tense using phonological mechanisms?

The rhymes of [brɪŋ] and [baɪ] don’t share any features with [ɔt] beyond [consonantal]. If we assume a floating pair of segments, /ɔt/, they can do so correctly and replace the root segments.

<table>
<thead>
<tr>
<th>bat</th>
<th>d, ɔt</th>
<th>MAX(float)</th>
<th>MAX(root)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>bat</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>bat</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>c.</td>
<td>bat</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>d.</td>
<td>baɪd</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cloning MAX(float) or MAX(root) will give a small probability to \(ɔt\)-taking, but will say nothing about the possible shapes of \(ɔt\)-takers.

The fact that the regular [baɪd] harmonically bounds the intended winner is also a hint that something non-phonological is going on, prompting the speaker to assume suppletion or some phonology-free diacritic.

Either cloning MAX(float) or using diacritics is equally bad for finding out what kind of roots are \(ɔt\)-takers, and indeed speakers have no clue about \(ɔt\)-taking.

5 Conclusions

Speakers learn statistical trends in their lexicon, and they do so in terms of UG. Now we have two ways of studying UG: Study regular phonology typologically, and study irregular morpho-phonology in individual languages.

To make sure that the grammar gets to see lexical statistics, don’t bury them in the lexicon, and work “inside-out”:

- Assume the paradigm’s base as the UR, derive the other forms from it.
- Assume that affixes only have segments in them, and try to get the rest from constraint interactions. Clone constraints as necessary.
- If no grammar can be found, assume that missing structure is floating in the UR’s of affixes, and try to get the rest from the grammar.
- If everything else fails, assume suppletion and/or diacritics.

This approach learns lexical trends and projects them onto novel words using an Optimality Theoretic grammar.

References


Robert Kirchner, Wolf Wikeley & Joe Pater (eds.) *Papers in Experimental and Theoretical
Linguistics*, University of Alberta, vol. 6. 24–35.
Ernestus, Miriam & R. Harald Baayen (2003). Predicting the Unpredictable: Interpreting
Hayes, Bruce (1995). On what to teach the undergraduates: Some changing orthodoxies
in phonological theory. In Ik-Hwan Lee (ed.) *Linguistics in the Morning Calm 3*, Seoul:
Hanshin. 59–77.
Hayes, Bruce (1998). On the richness of paradigms, and the insufficiency of underlying
representations in accounting for them. Handout for talk at Stanford.
In Ben Hermans & Marc van Oostendorp (eds.) *The derivational residue in phonology*,
Amsterdam: Benjamins. 175–205.
Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe (to appear). Natural and unnatural
constraints in hungarian vowel harmony. *Language*.
Inkelas, Sharon, Aylin Kuntay, John Lowe, Orhan Orgun & Ronald Sprouse (2000). Turkish
Inkelas, Sharon & Cemil Orhan Orgun (1995). Level ordering and economy in the lexical
phonology of Turkish. *Language* 71. 763–793.
Inkelas, Sharon, Cemil Orhan Orgun & Cheryl Zoll (1997). The implications of lexical
exceptions for the nature of the grammar. In Iggy Roca (ed.) *Derivations and Constraints
Kaisse, Ellen (1986). Locating Turkish devoicing. In M. Dalrymple et al. (ed.) *WCCFL 5*,
*Folia Linguistica 38: Special Issue on Voice*, Berlin: Mouton de Gruyter. 117–143.
Kenstowicz, Michael & Charles Kisseberth (1979). *Generative Phonology: Description and
*Phonology* 18. 81–108.
University Press.
dissertation, Rutgers University.
indexation. In Leah Bateman & Adam Werle (eds.) *UMOP: Papers in Optimality Theory III*,
Amherst, MA: GLSA. 1–36.
Pater, Joe (2009). Morpheme-specific phonology: Constraint indexation and inconsistency
resolution. In Steve Parker (ed.) *Phonological Argumentation: Essays on Evidence and
Motivation*, Equinox. 1–33.
and aspiration: Evidence from Russian, Hungarian, German, Swedish, and Turkish. *The
Linguistic Review* 23. 1–35.
Tesar, Bruce (1998). Using the mutual inconsistency of structural descriptions to overcome
Amherst, MA: GLSA, 469–483.
Tesar, Bruce & Alan Prince (2006). Using phonotactics to learn phonological alternations. In