Learning hidden structure in morphological bases

Highlights:

• I show that the traditional generative analysis, which attributes hidden structure to roots, makes the wrong predictions about statistical knowledge that speakers have.

• I propose a learning model that attributes hidden properties to constraint rankings, and if necessary, also to the UR’s of affixes.Attributing hidden structure to roots is done only as a last resort, via suppletion.

• My proposal makes OT-based work, which benefits from UG effects, compatible with assuming surface-true forms as UR’s (Albright 2008a).

1 Turkish voicing alternations

1.1 Grammar-based analysis

(1) bare stem possessive
   sop sop-u ‘clan’
   dʒop dʒɔb-u ‘nightstick’

My analysis: irregular intervocalic voicing

(2) The UR’s of [sop] and [dʒop] are /sop/ and /dʒop/
(3) The UR of the possessive is /u/ (actually just a high vowel)
(4) /sop + u/ → [sopu] requires IDENT(voice)-LAB ≫ *VTV
   /dʒop + u/ → [dʒɔbu] requires *VTV ≫ IDENT(voice)-LAB

The inconsistent ranking requirements trigger constraint cloning:

(5) IDENT(voice)-LAB_{sop} ≫ *VTV ≫ IDENT(voice)-LAB_{dʒop}

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

(6) /top + u/ IDENT(voice)-LAB *VTV
   a. ⋆ top-u
   b. tob-u *

(7) /ot + u/ IDENT(voice)-LAB *VTV
   a. ot-u *
   b. ⋆ od-u

(8) IDENT(voice)-LAB_{[sop, sop, alp, ...]} ≫ *VTV ≫ IDENT(voice)-LAB_{[dʒop, harp, ...]}

Until the speaker gets:

(9) IDENT(voice)-LAB_{22 items} ≫ *VTV ≫ IDENT(voice)-LAB_{8 items}

Novel p-final words will have a 8/30 (=27%) chance of alternating with [b]. The result: the lexical statistics are built into the grammar.

1.2 Why does this have anything do to with the grammar?

Becker, Ketrez & Nevins (2007) showed that Turkish speakers replicate the lexical statistics for nouns of different places (p, t, ʃ, k) and sizes (mono- vs. poly-syllabic), but do not replicate the lexical statistics about vowel height (more alternations after high vowels in the lexicon). We proposed that UG acts as a filter on the kinds of generalizations that speakers learn.

More generally, processes that are regular in some language are often irregular in another: intervocalic voicing, vowel harmony, cluster simplification, etc.

Using the same mechanism for regular and irregular processes seems like a good idea, especially given the dearth of regular processes.
1.3 What’s wrong with a UR-based analysis?

The classic generative analysis of Turkish (Inkelas & Orgun 1995; Inkelas et al. 1997):

(10) bare stem possessive

\[ \text{sop} \quad \text{sop-} \quad \text{‘clan’} \]
\[ \text{\&op} \quad \text{\&op-} \quad \text{‘nightstick’} \]

The analysis:

(11) The UR’s of \([\text{sop}]\) and \([\text{\&op}]\) are /sop/ and /\&op/.

(12) The UR of the possessive is /u/ (actually just a high vowel).

(13) /sop + u/ \rightarrow [sopu]

\[ \text{ IDENT(voice)-LAB} \gg *\text{VTV} \]

<table>
<thead>
<tr>
<th>sop + u</th>
<th>IDENT(voice)</th>
<th>*VTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. * sopu</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>b. sopu</td>
<td>*</td>
<td>!</td>
</tr>
</tbody>
</table>

(14) /\&op + u/ \rightarrow [\&opu] is consistent with IDENT(voice)-LAB \gg *VTV.

<table>
<thead>
<tr>
<th>&amp;op + u</th>
<th>IDENT(voice)</th>
<th>*VTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. * &amp;opu</td>
<td>*</td>
<td>!</td>
</tr>
<tr>
<td>b. * &amp;opu</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

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-height generalization that speakers don’t have.

Slightly better alternative that gets a consistent grammar: Attribute hidden structure of the affix.

(15) The UR’s of [sop] and [\&op] are /sop/ and /\&op/.

(16) The possessive has two allomorphs: /u/ and / [+voice] u/.

(17) /sop + u/ \rightarrow [sopu]

\[ /\&op + [+voice] u/ \rightarrow [\&opu] \]

The floating [+voice] is protected by \text{MAX(float)}, as in Wolf (2007), and we get a consistent grammar:

(18) \text{MAX(float)} \gg \text{IDENT(voice)-LAB}

Each allomorph of the possessive lists the roots it takes:

(19) /u/ takes /sop/, /tup/, /alp/, ...

\[ /[+voice] u/ \quad \text{takes} /\&op/, /harp/, ... \]

The prediction: Speakers will know the relative frequency of voicing alternations for the language as a whole, but not for specific stops or sizes, since the allomorphs of the possessive say nothing about the shape of the nouns they take.

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.
2 Fallback: When the grammar is not enough

Korean (Albright 2008b):

(20) Unmarked Accusative

nat’ natʰ il ‘piece’ 113
nat’ naṭʰ il ‘face’ 160
nat’ naḍil ‘grain’ 1
nat’ naṣil ‘daytime’ 17
nat’ nasil ‘sickle’ 375

Assuming /nat’/ for the roots and /il/ for the accusative can do some work:

(21)

<table>
<thead>
<tr>
<th>/nat’+ il/</th>
<th>*VTV</th>
<th>IDENT(voice)</th>
<th>IDENT(asp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.  natil</td>
<td>⋆</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.  nadil</td>
<td>⋆</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.  natʰ il</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(22) /nat’+ il/ → [natʰ il], [naṭʰ il]

requires *VTV ≫ IDENT(voice) ≫ IDENT(asp)

(23) /nat’+ il/ → [nadil], [naṭil]

requires *VTV ≫ IDENT(asp) ≫ IDENT(voice)

(24) IDENT(voice)_{113+160 items} ≫ IDENT(asp) ≫ IDENT(voice)_{1+17 items}

The prediction for a novel form, [pat’]:

(25) 94% chance of [tʰ], [ṭʰ], 6% chance of [d], [ḍ]

*Tf, which wants assimilation before a high vowel, will take care of [s]:

(26) /nat’+ il/ → [nasil]

requires *TI ≫ IDENT(cont)

(27) /nat’+ il/ → [natʰ il], [naṭʰ il], [nadil], [naḍil]

requires IDENT(cont) ≫ *TI

(28) IDENT(cont)_{113+160+1+17 items} ≫ *TI ≫ IDENT(cont)_{375 items}

The prediction for a novel form, [pat’]:

(29) 56% chance of [s], 44% chance of [tʰ], [ṭʰ], [d], [ḍ]

But are there plausible constraints that will map /nat’+ il/ to [naḍil] or [naṭʰ il]? It seems awfully hard to palatalize without a front vowel around.

If the speaker can’t find any such constraints, they will assume that the missing feature is floating in the UR of the accusative affix: /[−ant] il/.

(30) /nat’+ [−ant] il/ → [natʰ il], [naḍil]

requires MAX(float) ≫ IDENT(ant)

(31) /nat’+ [−ant] il/ → [naṭʰ il], [nadil]

requires IDENT(ant) ≫ MAX(float)

(32) /nat’+ [−ant] il/ → [nasil]

requires *f ≫ IDENT(ant), MAX(float)

(33) *f ≫ IDENT(ant)_{113+1 items} ≫ MAX(float) ≫ IDENT(ant)_{160+17 items}

The prediction for a novel form, [pat’]:

(34) 61% chance of [ṭʰ], [ṛ], 39% chance of [tʰ], [d]

Summary of the preferences that the grammar makes:

<table>
<thead>
<tr>
<th></th>
<th>IDENT(cont)</th>
<th>IDENT(voice)</th>
<th>IDENT(ant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>56%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[ṭʰ]</td>
<td></td>
<td>94%</td>
<td>61%</td>
</tr>
<tr>
<td>[tʰ]</td>
<td>44%</td>
<td>39%</td>
<td>16%</td>
</tr>
<tr>
<td>[ṛ]</td>
<td>6%</td>
<td>61%</td>
<td>2%</td>
</tr>
<tr>
<td>[d]</td>
<td>39%</td>
<td></td>
<td>1%</td>
</tr>
</tbody>
</table>

The high probability of [s] and [ṭʰ] conforms with the report of Albright (2008b). The probability of [tʰ] might be a bit too high.

5
3 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 *at-takers: teach, catch, think, bring, seek, fight, buy — how many of those can map to their past tense using phonological mechanisms?

The rhymes of [brnj] and [bAy] don’t share any features with [Ot] beyond [consonantal]. If we assume a floating pair of segments, /Ot/, they can dock correctly and replace the root segments.

(36)

\[
\begin{array}{c|c|c}
\text{buy} + \{d, \text{Ot}\} & \text{MAX(float)} & \text{MAX(root)} \\
\hline
а. & \text{hot} & * & ** \\
b. & \text{but} & * & * \\
c. & \text{buy} & ** \\
d. & \text{bAyd} & \\
\end{array}
\]

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

The fact that the regular [bAyd] 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 *at-takers, and indeed speakers have no clue about *at-taking.

4 Conclusions

Render onto the grammar what is the grammar’s.

- When faced with pairs of words in paradigms, assume one form as the UR and derive the other one from it.
- Assume that affixes only have segments in them, and try to get the rest from constraint rankings.
- 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 diacritic.

This approach learns lexical trends and projects them onto novel words without giving up the strengths of Optimality Theory.

References


