Michael Becker, UMass Amherst
michael@linguist.umass.edu

MUMM 2 @ MIT
March 29, 2008

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'
ʤop	ʤob-u	'nightstick'

My analysis: irregular intervocalic voicing

(2) The UR's of [sop] and [ʤop] are /sop/ and /ʤop/

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

(4) /sop + u/ → [sopu] requires IDENT(voice)-LAB ≫ *VTV
/ʤop + u/ → [ʤobu] requires *VTV ≫ IDENT(voice)-LAB

*Ideas presented today owe much to discussions with Adam Albright and Matt Wolf. I am also grateful to John McCarthy and Joe Pater for being a constant source of feedback, encouragement, and hard questions. I assume the responsibility for any remaining errors, in this paper and elsewhere.

The inconsistent ranking requirements trigger constraint cloning:

(5) IDENT(voice)-LAB$_{sop}$ ≫ *VTV ≫ IDENT(voice)-LAB$_{ʤop}$

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$_{\{sop,\ top,\ alp,\ ...\}}$ ≫ *VTV ≫ IDENT(voice)-LAB$_{\{ʤop,\ harp,\ ...\}}$

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, ʧ, 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	
sop	sop-u	'clan'
ʤop	ʤob-u	'nightstick'

The analysis:

(11) The UR's of [sop] and [ʤop] are /sop/ and /ʤoB/

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

(13) /sop + u/ → [sopu] requires IDENT(voice)-LAB ≫ *VTV

sop + u	IDENT(voice)	*VTV
a. ☞ sopu		*
b. sobu	*!	

(14) /ʤoB + u/ → [ʤobu] is consistent with IDENT(voice)-LAB ≫ *VTV

ʤoB + u	IDENT(voice)	*VTV
a. ʤopu		*!
b. ☞ ʤobu		

The grammar is consistent: IDENT(voice)-LAB ≫ *VTV

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 [ʤop] are /sop/ and /ʤop/

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

(17) /sop + u/ → [sopu]
/ʤop + [+voice] u/ → [ʤobu]

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

(18) MAX(float) ≫ IDENT(voice)-LAB

Each allomorph of the possessive lists the roots it takes:

(19) /u/ takes /sop/, /tup/, /alp/, ...
/[+voice] u/ takes /ʤop/, /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		
natˀ	natʰil	‘piece’	113
natˀ	natʃʰil	‘face’	160
natˀ	nadil	‘grain’	1
natˀ	naʤil	‘daytime’	17
natˀ	nasil	‘sickle’	375

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

(21)

/natˀ+ il/	*VTV	IDENT(voice)	IDENT(asp)
a. natil	*!		
b. nadil		*!	
c. ☞ natʰil			*

(22) /natˀ+ il/ → [natʰil], [natʃʰil]
requires *VTV ≫ IDENT(voice) ≫ IDENT(asp)

(23) /natˀ+ il/ → [nadil], [naʤil]
requires *VTV ≫ IDENT(asp) ≫ IDENT(voice)

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

The prediction for a novel form, [patˀ]:

(25) 94% chance of [tʰ], [tʃʰ], 6% chance of [d], [ʤ]

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

(26) /natˀ+ il/ → [nasil]
requires *TI ≫ IDENT(cont)

(27) /natˀ+ il/ → [natʰil], [natʃʰil], [nadil], [naʤil]
requires IDENT(cont) ≫ *TI

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

The prediction for a novel form, [patˀ]:

(29) 56% chance of [s], 44% chance of [tʰ], [tʃʰ], [d], [ʤ]

But are there plausible constraints that will map /natˀ+ il/ to [naʤil] or [natʃʰil]? 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) /natˀ+ [−ant] il/ → [natʃʰil], [naʤil]
requires MAX(float) ≫ IDENT(ant)

(31) /natˀ+ [−ant] il/ → [natʰil], [nadil]
requires IDENT(ant) ≫ MAX(float)

(32) /natˀ+ [−ant] il/ → [nasil]
requires *ʃ ≫ IDENT(ant), MAX(float)

(33) *ʃ ≫ IDENT(ant)$_{\{113+1\text{ items}\}}$ ≫ MAX(float) ≫ IDENT(ant)$_{\{160+17\text{ items}\}}$

The prediction for a novel form, [patˀ]:

(34) 61% chance of [tʃʰ],[ʤ], 39% chance of [tʰ], [d]

Summary of the preferences that the grammar makes:

(35)

	IDENT(cont)	IDENT(voice)	IDENT(ant)	
[s]	56%			= 56%
[tʃʰ]	44%	94%	61%	= 25%
[tʰ]			39%	= 16%
[ʤ]		6%	61%	= 2%
[d]			39%	= 1%

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

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 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 [bɑy] don't share any features with [ɔt] beyond [consonantal]. If we assume a floating pair of segments, /ɔt/, they can dock correctly and replace the root segments.

(36)

	bɑy + {d, ɔt }	MAX(float)	MAX(root)
a. ☞	bɔt		**
b.	bɑt	*	*
c.	bɑy	**	
d.	bɑyd		

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 [bɑyd] 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.

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 diacritics.

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

References

Albright, Adam (2008a). A Restricted Model of UR Discovery: Evidence from Lakhota. Ms. MIT.

Albright, Adam (2008b). Explaining universal tendencies and language particulars in analogical change. In Jeff Good (ed.) *Language Universals and Language Change*, Oxford University Press.

Becker, Michael, Nihan Ketrez & Andrew Nevins (2007). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish devoicing neutralization. Ms. UMass Amherst.

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 in Phonology*, Oxford: Clarendon. 393–418.

Wolf, Matthew (2007). For an autosegmental theory of mutation. In Leah Bateman, Michael O'Keefe, Ehren Reilly & Adam Werle (eds.) *UMOP 32: Papers in Optimality Theory III*, Amherst, MA: GLSA. 315–404.