From the lexicon to a stochastic grammar

Michael Becker, michael@linguist.umass.edu
University of Massachusetts, Amherst

1. The problem: Getting from a lexicon to a grammar

Phonological processes that are restricted to certain lexical items typically apply stochastically to novel items.

The behavior of novel items reflects lexical trends (Hayes & Londe 2006, Albright & Hayes 2003, Zuraw 2000, and several others)

→ **We need a way to project a stochastic grammar from the lexicon**

GLA (Boersma 1997):
The GLA doesn't distinguish existing words from novel words.
The GLA can't use a lexicon to learn a stochastic grammar, as it will wrongly promote general faithfulness (Hayes & Londe 2006, Tessier 2007).

USELISTED (Zuraw 2000):
Distinguishes existing items from novel ones, but doesn't derive the patterning of novel items from the trend created by the listed items.

2. Case study: Turkish voicing alternations

Turkish has intervocalic voicing in some words, but not others:

	ta**t**	ta**d**-ı	'taste'
	kana**t**	kana**d**-ı	'wing'
vs.			
	du**t**	du**t**-u	'mulberry'
	sepe**t**	sepe**t**-i	'basket'

Factors that correlate with the relative proportion of alternating stops:

Size:	Monosyllables don't usually alternate, polysyllables usually do
Place:	final [t]'s don't usually alternate, final [p], [k] usually do

Lexical statistics and experimental results that confirm speakers' knowledge of the pattern are in Becker, Ketrez and Nevins (2007).

How do speakers learn the proportion of alternating stops for each size and place?

3. The solution: Generalized Cloning

When lexical items demand conflicting rankings, BCD (Prince & Tesar 1999) detects inconsistency and stalls:

	OO-Ident(voice)	*VTV
dut-u ~ dud-u	W	L
tad-ı ~ tat-ı	L	W

The Pater (2006) solution: Clone a constraint to resolve the inconsistency.
Generalized cloning: **All** clones are lexically specific.

	OO-Ident (voice)$_{\{dut\}}$	*VTV	OO-Ident (voice) $_{\{tat\}}$
dut-u ~ dud-u	W	L	
tad-ı ~ tat-ı		W	L

Result: A categorical grammar for listed lexical items:

$$\text{OO-Ident(voice)}_{\{dut,\ \ldots\}} \gg \text{*VTV} \gg \text{OO-Ident(voice)}_{\{tat,\ \ldots\}}$$

Generalization: Lexically-specific grammar → Stochatic grammar

$$\text{OO-Ident(voice)}_{60\%} \gg \text{*VTV} \gg \text{OO-Ident(voice)}_{40\%}$$

4. Generalized Cloning: Specific constraints first

Speakers keep track of monosyllables independently of polysyllables. Speakers can do so thanks to the existence of initial-syllable faithfulness.

OO-Ident(voice)$_{\sigma1}$ accounts for fewer lexical items, i.e. it is more specific:

	OO-Ident (voice)$_{\sigma1}$	OO-Ident (voice)	*VTV
dut-u ~ dud-u	W	W	L
tad-ı ~ tat-ı	L	L	W
sepet-i ~ seped-i		W	L
kanad-ı ~ kanat-ı		L	W

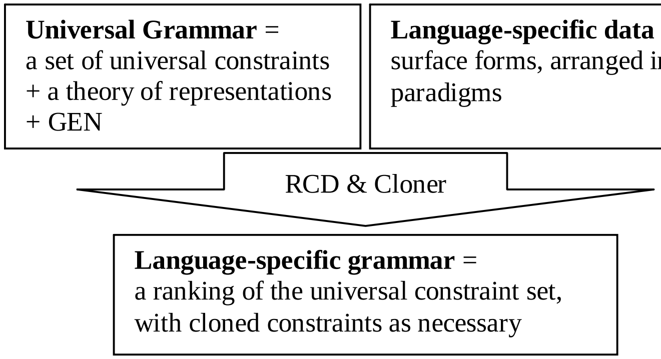
If the learner wrongly clones the general OO-Ident(voice) first, the general constraint will account for all exceptions, and the size effect will not be learned:

☹ OO-Ident(voice)$_{\{\text{dut, sepet, …}\}}$ » *VTV » OO-Ident(voice)$_{\{\text{tat, kanat, …}\}}$

The learner must clone OO-Ident(voice)$_{\sigma1}$ first to list monosyllables, then clone the general OO-Ident(voice) to list polysyllables:

OO-Ident(voice)$_{\sigma1\{\text{dut, …}\}}$,OO-Ident(voice)$_{\{\text{sepet, …}\}}$ » *VTV »
OO-Ident(voice)$_{\sigma1\{\text{tat, …}\}}$, OO-Ident(voice)$_{\{\text{kanat, …}\}}$

5. The learner



Universal Grammar =
a set of universal constraints
+ a theory of representations
+ GEN

Language-specific data =
surface forms, arranged into paradigms

RCD & Cloner

Language-specific grammar =
a ranking of the universal constraint set, with cloned constraints as necessary

The learner reads in the words of the given language one by one, and runs them through the grammar, creating a candidate set according to principles of OT-CC (McCarthy 2007). If the winner is different from the surface form, a winner-loser pair is formed and submitted to the RCD algorithm.

If RCD detects inconsistency, the learner clones a constraint that assigns the non-zero minimum of both W's and L's to the set of inconsistent ERC's. This continues recursively, until the data becomes consistent, or can't be made consistent by cloning.

The resulting grammar is categorical relative to existing lexical items, but can apply stochastically to novel items:

OO-Ident(voice)$_{\sigma1\{\text{dut, …}\}}$,OO-Ident(voice)$_{\{\text{sepet, …}\}}$ » *VTV »
OO-Ident(voice)$_{\sigma1\{\text{tat, …}\}}$, OO-Ident(voice)$_{\{\text{kanat, …}\}}$

The program uses code from JavaTableau (Becker, Potts & Pater 2007), CCamelOT (Becker 2005) and OT-Help (Becker, Pater & Potts 2007).

6. References

Albright, Adam and Bruce Hayes (2003) *Learning non-local environments*. Talk given at the LSA yearly meeting in Atlanta, Jan 4.

Becker, Michael, Nihan Ketrez and Andrew Nevins (2007) *The Surfeit of the Stimulus: analytic biases filter the statistics of Turkish voicing*. Ms.

Becker, Michael, Joe Pater and Christopher Potts (2007) *OT-Help: Typology explorer for Optimality Theory and Harmonic Grammar*. Open-source software, UMass Amherst. http://web.linguist.umass.edu/~OTHelp/

Boersma, Paul (1997). *How we learn variation, optionality, and probability*. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 21. 43–58.

Hayes, Bruce and Zsuzsa Londe (2006) *Stochastic phonological knowledge: the case of Hungarian vowel harmony*. Phonology 23. 59–104

McCarthy, Jonh J. (2007) *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox.

Pater, Joe (to appear) *Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution*. In Steve Parker, (ed.) Phonological Argumentation: Essays on Evidence and Motivation. London: Equinox.

Prince, Alan and Bruce Tesar (1999) *Learning Phonotactic Distributions*. ROA-353.

Tessier, Anne-Michelle (2007) *Biases and stages in phonological acquisition*. Ph.D. dissertation, University of Massachusetts, Amherst.

Zuraw, Kie (2000). *Patterned Exceptions in Phonology*. Ph.D. dissertation, UCLA.