The role of statistical generalizations in the grammar[1]

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

Highlights:

- Speakers have detailed knowledge about their lexicon. I show that this knowledge is grammatical, or biased by UG.
- General-purpose learners fail to replicate human behavior – universal biases are necessary for learning the lexicon like humans do.
- Theories of grammar that relegate information to the UR underplay speakers' knowledge of their language.
- I propose an OT model that accounts for lexical variation and projects statistical generalizations from them
- I present currently running experiments that aim to show the grammatical nature of lexical exceptions in Hebrew plural allomorphy

1 Turkish speakers' knowledge of their lexicon

This section summarizes joint work with Andrew Nevins (Harvard) and Nihan Ketrez (Yale). A manuscript is available upon request.

1.1 The generative approach to the lexicon

Turkish regulates the voicing of stem-final stops, productively enforcing final devoicing and intervocalic voicing (Lees 1961, Inkelas & Orgun 1995, Vaux 2005, and others):

(1)	rop ~ rob-u	< French [rɔb]	'dress'
	tüp ~ tüb-ü	< French [tüb]	'tube'
	gurup ~ gurub-u	< French [grup]	'group'
	ešarp ~ ešarb-ı	< French [ešarp]	'scarf'

But lexical exceptions abound:

(2)	Failure of final devoicing:	ad ~ ad-ı	'name'
	Failure of intervocalic voicing:	top ~ top-u	'ball'

Exceptions to final devoicing are fairly rare (~2% of the lexicon at best).
We focused on the application of intervocalic voicing, which affects ~54% of the lexicon.

Traditionally, generative linguists would derive the difference between *rop ~ rob-u* and *top ~ top-u* from a difference in the underlying representation:

(3)	Surface	UR
	rop ~ rob-u	/rob/ or /roB/
	top ~ top-u	/top/

Similarly, in OT terms (Prince and Smolensky 1993), Turkish is a final devoicing language, with the ranking IDENT(voice)$_{ONSET}$ » *VOICE » IDENT(voice):

/rob/	IDENT(voice)$_{ONSET}$	*VOICE	IDENT(voice)
a. rob		*!	
☞ b. rop			*

/rob-u/	IDENT(voice)$_{ONSET}$	*VOICE	IDENT(voice)
☞ a. rob-u		*	
b. rop-u	*!		*

/top-u/	IDENT(voice)$_{ONSET}$	*VOICE	IDENT(voice)
a. tob-u	*!	*	*
☞ b. top-u			

Both approaches have the same problem: If the difference between *rop ~ rob-u* and *top ~ top-u* is stored in the UR, it is hard to imagine how or why speakers will know the relative frequency of *rop*-like items vs. *top*-like items.

[1] Thanks to Lyn Frazier, John McCarthy and Joe Pater for valuable feedback and discussion. I also owe a huge debt of gratitude to Ram Frost of the psychology department at the Hebrew University. Remaining errors, if any, are my own

1.2 Statistical generalizations in the Turkish lexicon

Application of intervocalic voicing is unpredictable for any given existing lexical item, but certain factors are known to affect it. We mined an electronic lexicon (TELL, Inkelas et al. 2000) for such factors:

(4) **Size**

	n	% alternating
Monosyllabic, simplex coda (CVC)	137	12%
Monosyllabic, complex coda (CVCC)	164	26%
Poly-syllabic (CV.CVC and longer)	2701	59%

(5) **Place** of articulation of the stem-final stop

	n	% alternating
p	294	84%
t	1255	17%
č	191	61%
k	1262	85%

(6) **Height** of the stem's final vowel:

	n	% alternating
–high (a, e, o ,ö)	1690	42%
+high (ɨ, i, u, ü)	1312	72%

(7) **Backness** of the stem's final vowel:

	n	–back	+back	difference
p	294	90%	79%	–11%
t	1255	16%	18%	+1%
č	**191**	**44%**	**74%**	**+30%**
k	1262	84%	86%	+1%

What do speakers know about these numbers?
Inkelas & Orgun (1995) and Inkelas et al. (1997) mention **size** and **place**, but no vowel effects.

1.3 Experiment: Speakers' knowledge about voicing alternations

In our experiments, speakers replicate the **size** and **place** effect. Speakers do not replicate any vowel effects – neither **height** not **backness**.

We showed 24 adults novel nouns, e.g. *köč*. They were shown a possessor, and asked to choose between two vocal renditions of the possessed noun: *köč-ü* or *köǰ-ü*.

(8) Sizes: CVC, CVCC and CVCVC
Places: p, t, č, k
Vowels: a, ɨ, e, i, o, u, ö, ü (high, back, round)
Total 72 stimuli

(9) Logistic regressions on the lexicon and on the experiments results:

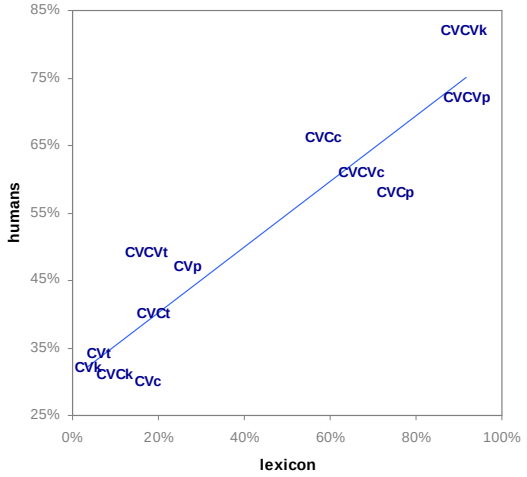
	Lexicon	Experiment
size	**p < .001**	**p < .001**
place	**p < .001**	**p = .005**
high	**p < .001**	ns
back	**p < .001**	ns
round	ns	ns

High and **back** are significant in the lexicon, not in speakers' choices.

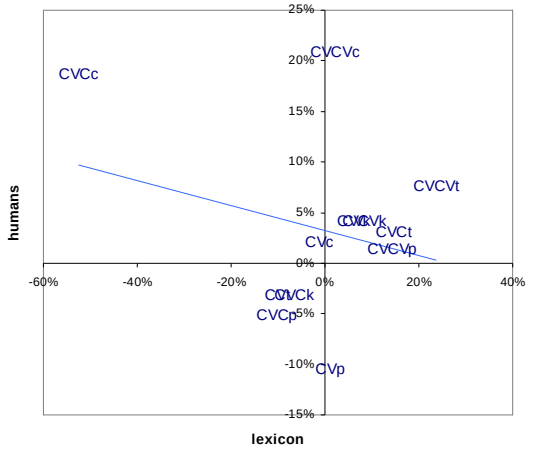
(10) ANOVA on the experimental results (*n*=24):

size	p < .001
place	p = .006
high	ns
back	ns

(11) The size and place effects in the lexicon and in the experiment



(12) The height effect in the lexicon and in the experiment, categorized by size and place. Positive values indicate more alternations with +high vowels, negative values indicate more alternations with –high vowels.



1.4 UG-less modeling of speakers' knowledge

Speakers' knowledge is tightly correlated with the lexicon; we assume that they project their knowledge from the lexicon. We used the Minimal Generalization Learner (henceforth MGL, Albright & Hayes 2002), which can learn generalizations from a corpus and project them onto novel words.

We gave the MGL the stop-final words of TELL and their possessive forms, annotated for mono-/poly-syllabicity of the stem, and a table of segments and their features. The MGL starts with creating a rule for each word of the lexicon, and collapses phonologically similar rules to form generalizations. Ex.:

(13)

šarap$_2$ ~ šarab**i**$_2$	rule:	p → b**i** / š a r a __ $_2$
kebap$_2$ ~ kebab**i**$_2$	rule:	p → b**i** / k e b a __ $_2$
Generalization:		p → b**i** / X a __ $_2$

(14) To make the MGL results comparable to the experimental results, we simulated 24 "learners", mapping the MGL results (with some random noise) onto zeros and ones.

(15) Logistic regressions on the lexicon, on the experimental results, and on the MGL results:

	Lexicon	MGL	Experiment 2
size	**p < .001**	**p < .001**	**p < .001**
place	**p < .001**	**p < .001**	**p = .005**
high	**p < .001**	**p = .007**	ns
back	**p < .001**	**p = .003**	ns
round	ns	ns	ns

The MGL replicates the **high** and **back** effects of the lexicon, unlike the humans.

The MGL modeled the lexicon *too well*, generating data that fit the lexicon more closely than humans. Most importantly, it failed to ignore the vowel effects that are in the lexicon.

We conclude that the MGL reproduced both the phonologically-motivated generalizations (**size** and **place** effects) and the accidental generalizations (**high** and **back** effects) that were found in the lexicon.

What's missing from the MGL is a theory of possible and impossible interactions between phonological elements.

2 The ingredients of a UG-based analysis

I assume that UG acts as a filter on learning the lexicon. UG constrains the learning process, making speakers notice phonologically-motivated generalizations and ignore others.

When speakers derive novel forms, they **do not access their lexicon**. They only use their grammar, which has the phonologically-motivated aspects of the lexicon built into it.

2.1 Lexical statistics are kept following Inconsistency Detection

I propose a learning model in which speakers detect inconsistency in the grammar (Pater 2006) and then start keeping track of the behavior of individual items:

(16)

/rop+u/	OO-IDENT(voice)	*VT]V
→ rob-u	1	
rop-u	L_0	W_1

(17)

/top+u/	OO-IDENT(voice)	*VT]V
→ top-u		1
tob-u	W_1	L_0

(18) OO-IDENT(voice)$_{top}$ » *VT]V » OO-IDENT(voice)$_{rop}$

As more words are learned, each instance of OO-IDENT(voice) will accumulate "weight", and this "weight" is projected onto novel words:

(19) OO-IDENT(voice)$_{top,ip,sop,bap,čap,hep,kip…}$ » *VT]V » OO-IDENT(voice)$_{rop,tüp,kap…}$

Thus, the ratio of alternating and non-alternating nouns is built into the constraint ranking. A novel word like *züp* will be attracted by the heavier top-ranking OO-IDENT, so *züp-ü* is more likely than *züb-ü*.

2.2 Generalizations in terms of constraints

Inconsistency Detection is done for each constraint in CON separately:

(20) The place effect follows from the existence of place-specific faithfulness constraints: OO-IDENT(voice)-COR, etc.

(21) Initial syllables are protected by positional faithfulness, allowing generalizations over mono-syllabic bases to be kept separately from generalizations over poly-syllabic bases: OO-IDENT(voice)-COR$_{\sigma 1}$, OO-IDENT(voice)-LAB$_{\sigma 1}$, etc.

2.3 Lack of effect from lack of constraints

No language is known to change obstruent voicing based on the quality of a neighboring vowel.

Therefore, OT has no constraints relating any voice specification and any neighboring vowel quality, such as:

(22) *[+back][–voice]

Any relationship between vowel quality and obstruent voicing is necessarily accidental; speakers cannot be attuned to it.

3 The grammar of Hebrew plural allomorphy

I am greatly indebted to Ram Frost of the psychology department at the Hebrew University, who generously offered to run my experiments at his laboratory for verbal information processing.

3.1 The Lexicon

Hebrew has two plural suffixes, *–im* and *–ot*, with a partially-predictable distribution. Above the word level, completely predictable gender agreement reveals that *–im* is masculine and *–ot* is feminine. At the word level, native[2] nouns can take a mismatching suffix:

(23)
xalon-ót	gdol-ím	
window-pl	big-pl	'big windows'
nemal-ím	ktan-ót	
ant-pl	small-pl	'small ants'

In the loanword phonology, the plural suffix selection is completely regular even at the word level: If the right edge of the word is recognizable as a feminine suffix, *–ot* is selected, otherwise it's *–im*.

(24)
artišók	artišók-im	* artišók-ot	'artichoke'
kolég-a	* kolég-im	kolég-ot	'colleague'
madám	? madám-im	??? madám-ot	'madam (in a brothel)'

My data comes from an electronic dictionary (Bolozky & Becker 2006). Native masculine nouns that take *–ot* are more common than native feminine nouns that take *–im*:

(25)
	-im	-ot
masculine	3716	173
feminine	23	1196

[2] I use the term "native" as a label for a synchronically-defined class of nouns, ignoring etymology. Native nouns are characterized here by movement of stress to the plural suffixes. See Becker (2003) for other properties of native nouns in Hebrew.

The choice of plural suffix is not predictable from the singular form. There are even some minimal pairs:

(26)
himnon-ím / himnon-ót	'national anthem' / 'religious hymn'
tor-ím / tor-ót	'line, queue', 'appointment' / 'turn'
maamad-ím / maamad-ót	'stand' / 'status'
mazal-ím (tov-ím) / mazal-ót	'good luck' / 'astrological sign'

The label "unpredictable", however, misses the partial phonological predictability of the suffix for masculine native nouns:

(27)
	–im	*–ot*	
a	1136	37	3%
e	788	26	3%
i	422	9	2%
o	300	**96**	**24%**
u	1070	5	<1%
Total	3716	173	

Berent, Pinker & Shimron (1999) show that speakers project this trend onto novel items, choosing *–ot* more often with nouns that have [o] in their final syllable.

3.2 The role of markedness - Universality

In OT, markedness constraints have three properties:
a. They are universal (and possibly innate)
b. Their effect is general by default
c. They assess output forms only

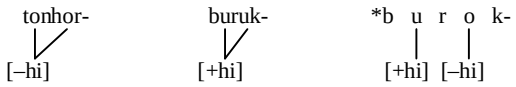
If Hebrew exceptions are organized using universal constraints, we expect to see the ***exceptional*** Hebrew pattern as a ***regular*** pattern in some other language.

In Shona (Beckman 2004), mid vowels (e, o) are licensed in initial syllables, or adjacent to another mid vowel:

(28) tonhor-, bover-, verer-, pofomar-
buruk-, simuk-, kumbir-, katuk-
*burok-, *boruk-, *burek-

Beckman analyzes the pattern using IDENT(high)$_{\sigma 1}$ » *MID » IDENT(high)

(29)



In Shona, the strong position that licenses mid vowels is the initial syllable. In Hebrew, it is the stressed syllable. Licensing of mid vowels only in stressed syllables is very common, e.g. in Russian, many dialects of Arabic, Portuguese, and others. Hebrew allows the stressed mid vowel to license an adjacent mid vowel as well.

(30)

	singular	*plural*
irregular	xalon	x a l o n – ó t [–hi]
regular	alon	a l o n – í m [–hi] [+hi]

Unlike in Shona, in Hebrew alon-ím is grammatical even with its unlicensed mid vowel. Mid vowel licensing in Hebrew is emergent.

A small number of Hebrew nouns avoid an unlicensed mid vowel by changing the root vowel: *xók ~ xuk-ím* 'law', *néc ~ nic-ím* 'hawk'.

/alon$_m$/+{im$_{mpl}$, ot$_{fpl}$}	IDENT(hi)	AGREE(gender)	*MID
alon-im			*
alon-ot		*!	
alun-im	*!		

/xalon$_m$/+{im$_{mpl}$, ot$_{fpl}$}	IDENT(hi)	*MID	AGREE(gender)
xalon-im		*!	
xalon-ot			*
xalun-im	*!		

/xok$_m$/+{im$_{mpl}$, ot$_{fpl}$}	*MID	AGREE(gender)	IDENT(hi)`
xok-im	*!		
xok-ot		*!	
xuk-im			*

3.3 The role of markedness – generality

In the lexicon, feminine nouns overwhelmingly take *–ot* regardless of the stem's vowel. In fact, the only [o]-final noun takes *–im* rather than *–ot*, so if anything, there is a slight preference for *–im*, not *–ot*,after [o].

(31)

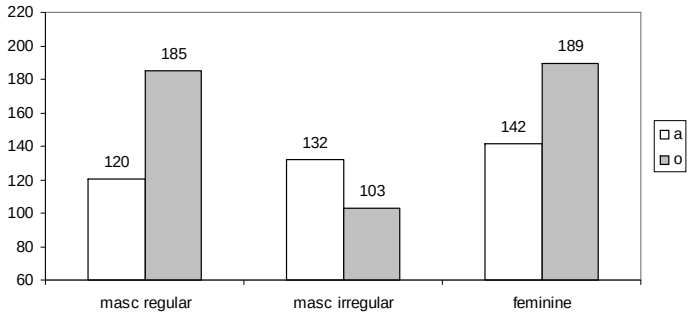
	–im		*–ot*
aa	3	2%	124
ea	1	2%	62
ia	7	2%	453
oa	2	5%	42
e	3	75%	1
ee	5	16%	26
i	1	50%	1
io	1	100%	0
Total	23		709

Experiment in the works: lexical decision task on actual roots with either suffix:

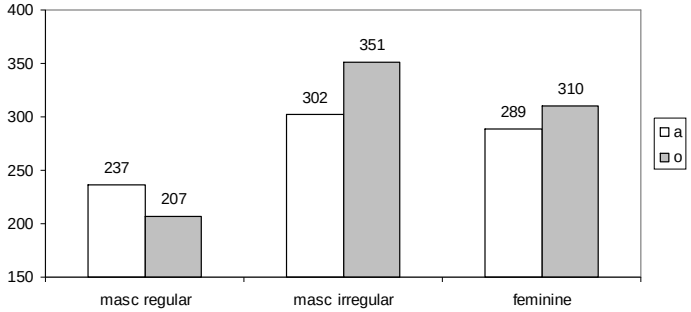
(32)

masculine, regular		masculine, irregular		feminine, regular	
[a]	[o]	[a]	[o]	[a]	[o]
agasím	aloním	gagót	aronót	agadót	agorót
bcalím	egozím	ilanót	borót	avkót	bsorót
duxaním	egrofím	kravót	cinorót	kfafót	ofnót
kfarím	kiyorím	znavót	mekomót	taxanót	tyotót
*agasót	*alonót	*gagím	*aroním	*agadím	*agorím
*bcalót	*egozót	*ilaním	*borím	*avkím	*bsorím
*duxanót	*egrofót	*kravím	*cinorím	*kfafím	*ofním
*kfarót	*kiyorót	*znavím	*mekomím	*taxaním	*tyotím

(33) Real words



(34) Incorrect suffix



The vowel effect in the masculine nouns is expected, basically replicating the results from Berent, Pinker & Shimron (1999). The pleasant surprise is the vowel effect on the feminine nouns, since in the lexicon they overwhelmingly take *–ot*, regardless of the root's vowel.

A general-purpose learner should not produce a vowel effect in the feminines, since the in the lexicon, the vowel effect is limited to masculine nouns.

3.4 The role of markedness – assessing outputs

Experiment in the works: choosing plural suffixes with vowel alternations that are not attested in actual Hebrew.

		mapping	training	novel items
Language A	a.	[ao] → [ai]	acok ~ acikot apoz ~ apizot abol ~ abilim azod ~ azidim	agof, ados, axos, amox, atox, alog, aroš, adoc
	b.	[aa] → [au]	amag ~ amugot afaš ~ afušot anar ~ anurim axac ~ axucim	axaf, ayav, apas, azax, abak, ataz, adal, ayad
Language B	a.	[ai] → [ao]	acik ~ acokot apiz ~ apozot abil ~ abolim azid ~ azodim	agif, adis, axis, amix, atix, alig, ariš, adic
	b.	[au] → [aa]	amug ~ amagot afuš ~ afašot anur ~ anarim axuc ~ axacim	axuf, ayuv, apus, azux, abuk, atuz, adul, ayud

Speakers learn novel names for common nouns (all fruits and vegetables whose Hebrew name is masculine and takes *–im*). They learn the singulars and the plurals, and then asked to supply plurals for new nouns.

My prediction: When deriving novel nouns, speakers will form a strategy for choosing the plural suffix based on the vowel of the stem (either in the input or in the ouput).

Initial pilot results are promising!

My success will be devastating for any general-purpose learner that simply learns the lexicon without universal biases. Since no Hebrew noun has either [i] → [o] or [o] → [i], no prediction at all is made about the choice of plural allomorph.

4 Conclusions

- Speakers use their universal grammar when they learn the words of their language.
- Lexical exceptions are learned in terms of rankings of universal constraints, and these rankings can be projected unto novel nouns.
- I am out to show that UG-less learning algorithms fail to model human behavior in a range of different ways.

References

Albright & Hayes (2002). Modeling English Past Tense Intuitions with Minimal Generalization. In Maxwell, Michael (ed) *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*. Philadelphia, July 2002. ACL.

Bolozky, Shmuel and Michael Becker (2006) Living Lexicon of Hebrew Nouns. Ms. UMass Amherst.

Inkelas, Sharon, Aylin Kuntay, John Lowe, Orhan Orgun & Ronald Sprouse (2000). Turkish electronic living lexicon (TELL). Website, http://socrates.berkeley.edu:7037/.

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.

Pater, Joe (2004). Exceptions and optimality theory: Typology and learnability. Talk given in the Conference on Redefining Elicitation: Novel Data in Phonological Theory. New York University.

Vaux, Bert (2005). Formal and empirical arguments for morpheme structure constraints. Talk given at LSA Annual Meeting, San Francisco.