Class 7: Do humans smooth?

To do

- Read Moreton & Pater 2012
- MaxEnt exercise with priors:
  - Run the "double-tweak" case from previous exercise, but this time in MaxEnt Grammar Tool (http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/)
  - Run once ignoring "open constraints" button (will use default values)
  - Run again making a constraints file—use SameConstraintFile.txt in the MaxEnt folder as a basis—but instead of 10,000 for σ^2, use something much smaller, like 0.1
  - Briefly compare and contrast results.

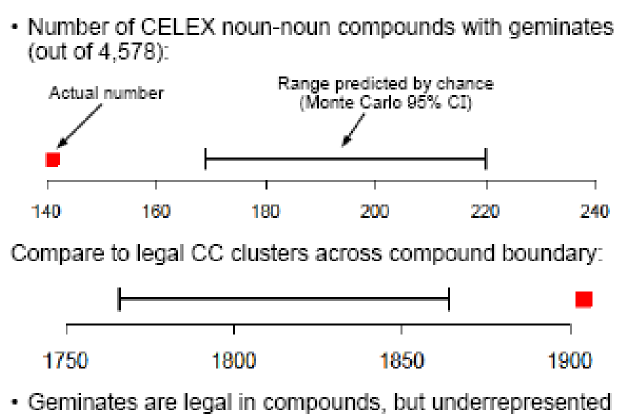
1 Smoothing bias

- We saw that smoothing (a.k.a. regularization) is a way to avoid overfitting:
  - Tell your software to find a model that compromises between fitting the data and staying close to default parameter values
- For regression coefficients or MaxEnt weights, typical default is 0 for everything
- This is all well and good for modeling, but do people do it when learning variation?
- That is, beyond any substantive biases (which Bruce will discuss Thurs.), do human learners have a "smoothing bias" to keep weights small?

Case study I: Martin 2007a , Martin 2011

2 Facts to be accounted for

- English does not allow geminates (long/double consonants) within a morpheme: there can be no minimal pair [hæpi]/[hæppi].
- English does allow geminates in compounds and affixed words: *no[nn]egotiable*, *sou[ll]ess*, *boo[kk]ase*.
- Martin discovered, however, that geminates are less common than would be expected by chance—that is, there are not as many words like *bookcase* as expected:



(Martin 2007b)

3 Martin's approach

- It's easy to construct a learner that can learn these facts.
- What Martin set out to do was construct a learner that, presented with no trend in compounds, will <u>learn to avoid geminates in compounds anyway</u>.

4 Martin's toy language—contains only 2 sounds

- The training data consists of biconsonantal clusters of [p] and [t], with an optional morpheme boundary:

Cluster	Structure	Number of examples
pt	monomorpheme	2000
tp	monomorpheme	2000
p+t	compound	1000
t+p	compound	1000
p+p	compound	1000
t+t	compound	1000

No bias in training data

- Tautomorphemic geminates [pp], [tt] do not occur in training data, but heteromorphemic geminates occur freely

(Martin 2007b)

5 Constraints available to learner

Structure-sensitive constraints:

***pp** no geminates within morpheme
***tp** no non-geminate clusters within morpheme
***p+p** no geminates across morpheme boundary
***t+p** no non-geminate clusters across morpheme boundary

Structure-blind constraints:

***p(+)p** no geminates
***t(+)p** no non-geminate clusters

(Martin 2007b)

6 MaxEnt Grammar learned (Martin 2007b version, since weights all non-negative)

	*pp weight = 4.01	*tp weight =0.13	*p(+)p weight =0.03	*t(+)p weight =0.00	*p+p weight =0.00	*t+p weight =0.00	score	probability
a pp	*		*				$e^{-4.04}=0.02$	1%
☞ *b* tp		*		*			$e^{-0.13}=0.87$	31%
c p+p			*		*		$e^{-0.04}=0.96$	34%
d t+p				*		*	$e^{-0.00}=1.00$	35%

- *pp* gets a low score, as expected—because *pp has a big weight
- *tp* gets a high score, as expected—because *tp has a small weight
- *t+p* gets a high score, as expected
- but *p+p* gets a slightly lower score—**because *p(+)p has a non-negligible weight**

7 Why does *p(+)p get non-zero weight?

- Recall the form of the Gaussian prior: the learning model is trying to maximize...

$$\ln(\textit{probability}(\text{data under model})) - \sum_{j=1}^{M} \frac{(w_j - \mu_j)^2}{2\sigma_j^{\,2}}$$

- Assume a μ of 0 for all constraints C_j
- The smoothing term uses $(w\text{-}0)^2 = \boldsymbol{w^2}$
  - So, it's better to account for data like the absence of *pp* by spreading the responsibility over two constraints—*pp and *p(+)p—than by loading all the blame onto one constraint. (Let's check the math)
- Thus, if there are structure-blind constraints like *p(+)p, generalizations that are true of one type of word (here, monomorphemes) will "leak" onto other types of word (here, compounds).

8 Similar finding in Navajo compounds

(Na-Dene language from the U.S. with about 149,000 speakers [(Lewis 2009)])

- Within a word, sibilants *must* agree—affixes even alternate:

/ **s**ì+**t͡ʃ**ìd /	→	**ʃ**ì+**t͡ʃ**ìd	'he is stooping over'
/ **s**ì+téː**ʒ** /	→	**ʃ**ì+téː**ʒ**	'they two are lying'
/ ji+**s**+léː**ʒ** /	→	ji+**ʃ**+t͡ɬéː**ʒ**	'it was painted'
/ ji+**s**+ti**z** /	→	ji+**s**+ti**z**	'it was spun'

- In compounds, they *tend* to agree (if in adjacent syllables)

70% agree:

ts^{h}eː + ts'in	'tailbone'
k'iːʃ + ʒin + iː	'blue beech'
ts^{h}é + zéí	'gravel'

30% disagree—by chance, you'd expect 37%-55%

tʃéí + ts'i:n	'rib cage'
ts^{h}é + tʃéːʔ	'amber'

9 Similar finding in Turkish compounds (Martin 2007a only)

- Vowels within a stem tend strongly to agree in backness
- Vowels within a lexicalized compound tend—less strongly but still significantly—to agree

"single-word" (lexicalized) compounds

baş + bakan	'prime minister'	60% agree
ön + ayak	'pioneer'	40% disagree (expect 44%-54%)

- Non-lexicalize ("izafet") compounds have *more* disharmony than expected
  - Martin speculates that this is because disharmonic compounds are less likely to get lexicalized (become single-word), and thus remain in the izafet class

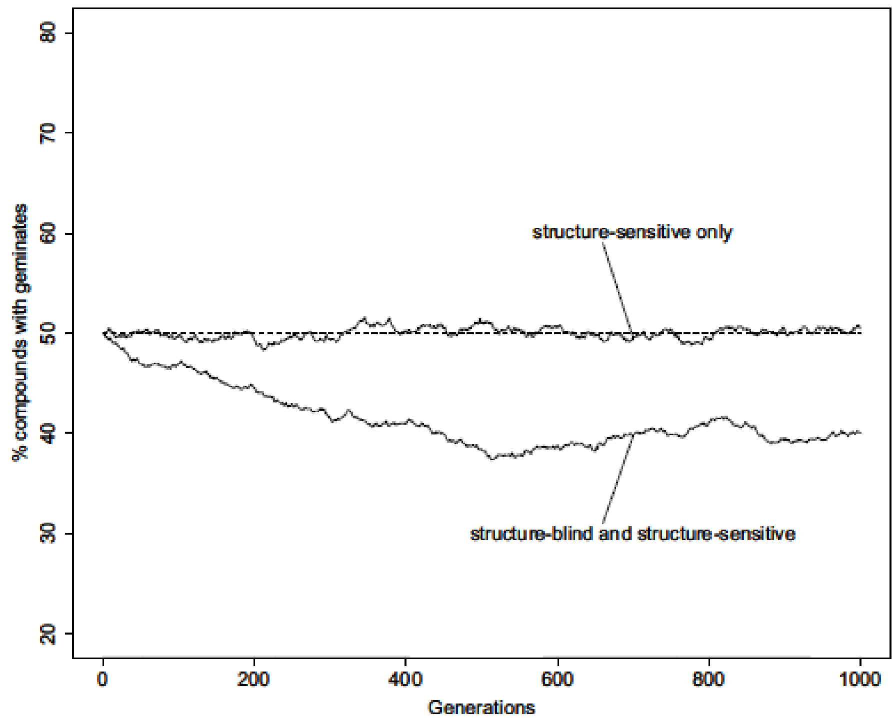
"izafet" (productive) compounds

baş + ağrı+sı	'headache'	48% agree
deniz + bız + ı	'mermaid'	52% disagree (expect 45%-50%)

10 Summary of Martin's argument

- Learners have available various versions of a markedness constraint: within-word, across-word, and unspecified
- If you train a learner on data where the constraint holds only within-word...
  - The Gaussian prior says it's better to blame both *PP and *P(+)P [contrary to evidence] rather than just *PP
- Those learners will slightly avoid compounds that violate the constraint
  - Their children now train on data where there's evidence for *P+P too

- Generation after generation, the avoidance grows:



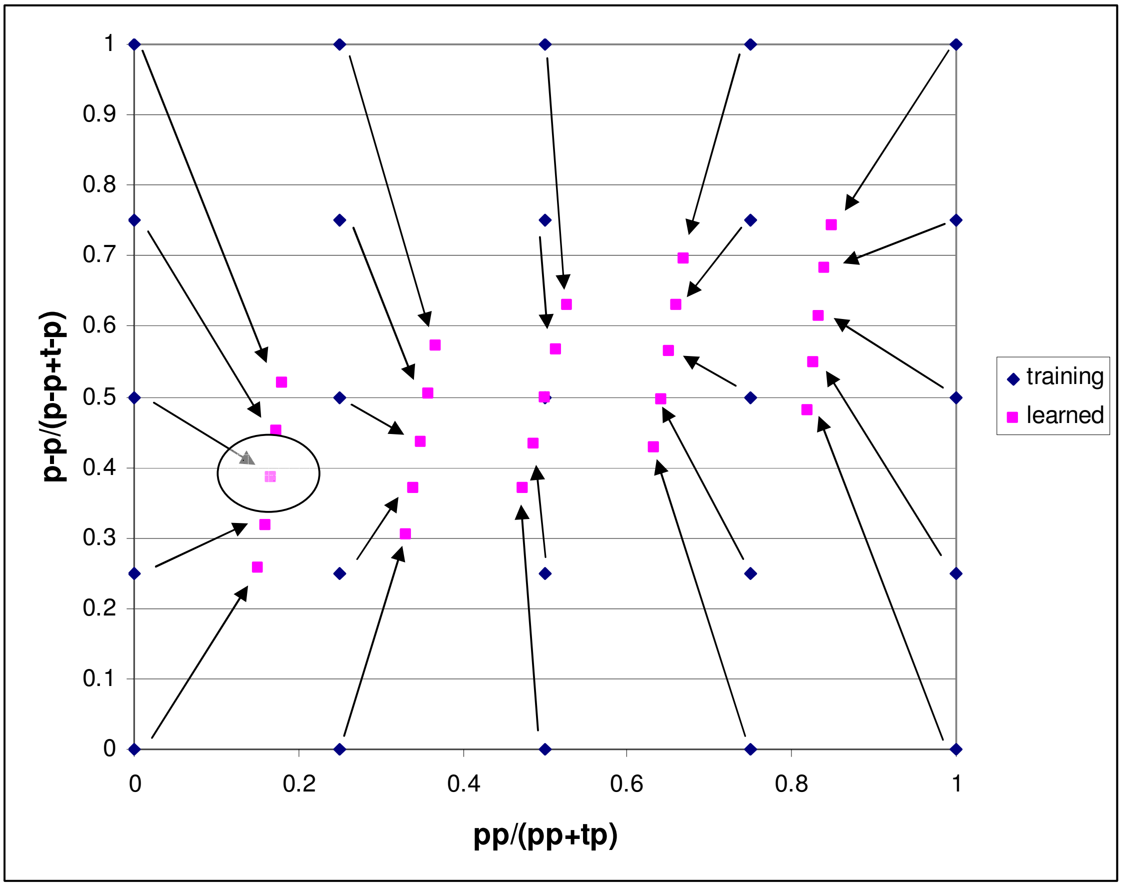
(Martin 2011, p. 765)

Discussion

11 How good are we at frequency matching?

- Suppose your hypothesis is: *Learners are very good at learning the degree to which geminates are dispreferred in compounds*
  - Discuss: in light of Martin's article, what should the null hypothesis be?

12 A little simulation



- Training files look like this (with freq. column varying)
- all μ=0, all σ^2=0.5 (very conservative)

			*pp	*tp	*p+p	*t+p	*p(+)p	*t(+)p
			*pp	*tp	*p+p	*t+p	*p(+)p	*t(+)p
in	pp	5	1				1	
	tp	15		1				1
	p+p	3			1		1	
	t+p	1				1		1

- Suppose learners (children, experimental subject) have been exposed to (0.18, 0.39) (circled)
  - and suppose they then show in some tasks a preference for (0.18, 0.39)
  - Is that because they learned what they were exposed to?
  - Or is it because they ignored much of the learning data, treating their input as (0, 0.5), and just smoothed it?

13 How good are we at frequency matching?

- The message I take from this work is that we may want to ask not
  - "Do speakers (or experimental subjects) demonstrate implicit knowledge of the details of the variation they've been exposed to"

  but
  - "Do they demonstrate such knowledge beyond or counter to what's expected from a rough grasp of the data and then smoothing (or other) bias?"

Case study II: Ryan 2010

14 Tagalog affix order

- A famous case of free variation that's morphological rather than strictly phonological
- CV- reduplication can mark incomplete aspect
- But its position seems to vary, with no apparent difference in meaning

  ma-ka-pag-**pa**-pa-bili ~ ma-**ka**-ka-pag-pa-bili 'will be able to have someone buy'
  ability-telic-transitive-**incomplete**-causative-BUY ~ abil-**incompl**-tel-trans-caus-BUY
  (p. 759)

- We'll skip over the substance of Ryan's analysis—a set of markedness constraints on affix order

15 Learning simulation

- Learning data limited to just the most-frequent candidate in each case
- Noisy Harmonic Grammar (no explicit smoothing term)
- If left to run long enough, the learner fits the (incorrect) training data
- But if stopped early—a form of smoothing—the learner predicts that other candidates get some probability to
  - Result: a good match to the actual (untrained) frequencies for each candidate

OUTPUT	ACTUAL CORPUS	SEEN BY LEARNER	GENERATED BY LEARNER
ma-RED-ka-pag-ROOT	61.8%	100%	69.2%
ma-ka-pag-RED-ROOT	38.2%	0%	29.3%
ma-RED-ki-pag-ROOT	99.9%	100%	97.1%
ma-ki-pag-RED-ROOT	0.1%	0%	2.9%

TABLE 5. A closer look at some of the learner's predictions.

(p. 774)

  - Conclusion: the speaker doesn't need to track detailed variation rates; just needs to note the main trends, and not fit too closely

What about the reverse bias?

16 Simplicity bias?

- A Gaussian prior likes to spread responsibility among multiple constraints
- Bruce pointed out that the opposite prior—dependent on square roots of weights, say—would be a "simplicity bias", in the sense that weight prefers to be heaped onto a small number of constraints
- Is there any evidence for this?

17 Phonologization, from Hayes 1999

- Many factors affect how much aerodynamics favors voicing vs. voicelessness (see Ohala 1983, Westbury & Keating 1986) (Hayes p. 8)
  - place of articulation: fronter closure → bigger oral chamber → more room for the air → airflow across glottis encouraged for longer
  - closure duration: as time passes during the closure, more air pressure in oral chamber → airflow across glottis discouraged
  - being after a nasal: nasal leak and velar pumping encourage airflow
  - being phrase/utterance-final: subglottal pressure is lower → airflow across glottis discouraged

- Hayes constructs the following "difficulty landscape" using an aerodynamic model (Keating 1984):
  - 0 means there's no problem having voicing; bigger numbers mean it's difficult.

(2) Landscape of Difficulty for Voiced Stops: Three Places, Four Environments

	b	d	g	
[-son] ___	43	50	52	
# ___	23	27	35	
[+son, -nas] ___	10	20	30	
[+nas] ___	0	0	0	contour line: 25

(p. 9)

- The thing is, there is no language that draws the line at 25. Instead, languages draw vertical or horizontal lines that partly contradict the phonetics:
  - *g (as in Dutch): ignores the fact that initial [g] is easier than post-obstruent [d]

- This can lead to seeming markedness contradictions in the corners:
  - *p (as in Arabic): even in geminates, you get only [bb], not *[pp]
  - *VOICEDGEMINATE (as in non-loan Japanese): only [pp], not *[bb]

18 Hayes's proposed solution

- The learner...
  - ...compiles a difficulty map like the above
  - ...constructs constraints according to templates (*[αF], *[αF][βG], *[αF,βG], etc.)
  - ..evaluates constraints according to how often they correctly predict that one item in the map is harder than another
    - e.g., *g: correct about g/[-son]__ vs. d/[-son]__, wrong about g/#__ vs. d/[-son]__
    - collect % of pairs for which prediction is correct
  - ...to be accepted, a constraint must do better on the above test than all its "neighbors" that are equally or less complex
    - constraints are neighbors if they differ in just one symbol (whatever counts as a symbol in your theory).
    - e.g., *[coronal, +voice] and *[dorsal, +voice] are neighbors, equally complex
    - *g and *#g are neighbors; *g is less complex than *#g
- Result: The learner adds complex constraints only if they justify themselves.

constraints like *[dorsal, +voice] and *[+nasal][-voice], but nothing more complex.

19 What kind of bias is this?

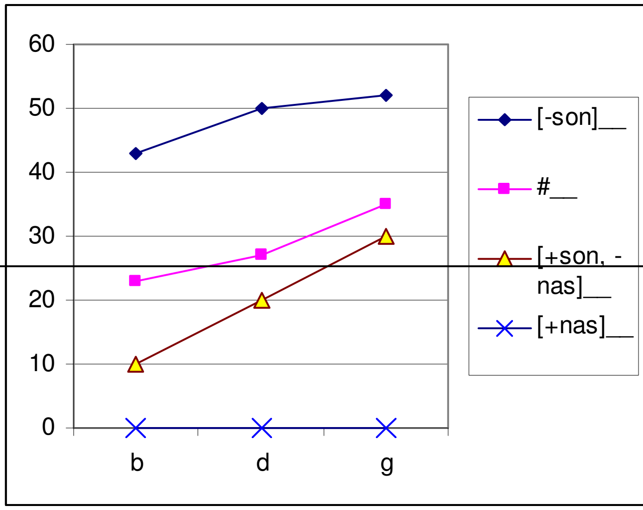
Simplicity or share-the-burden?

Suppose that the difficulty scores above were reflected in actual variation (training file below)—what will a learner draw from such data?

			“column” constraints			“row” constraints					“cell” constraints											
			*b	*d	*g	*[-son][+voice]	*#[+voice]	*[+son,-nas][+voice]	*[+nas][+voice]	*[-son]b	*#b	*[+son,-nas]b	*[+nas]b	*[-son]d	*#d	*[+son,-nas]d	*[+nas]d	*[-son]g	*#g	*[+son,-nas]g	*[+nas]g	Ident
[-son]b	b	57	1			1				1												
	p	43																				1
#b	b	77	1				1				1											
	p	23																				1
[+son,-nas]b	b	90	1					1				1										
	p	10																				1
[+nas]b	b	100	1						1				1									
	p	0																				1
[-son]d	d	50		1		1								1								
	t	50																				1
#d	d	73		1			1								1							
	t	27																				1
[+son,-nas]d	d	80		1				1								1						
	t	20																				1
[+nas]d	d	100		1					1								1					
	t	0																				1
[-son]g	g	48			1	1												1				
	k	52																				1
#g	g	65			1		1												1			
	k	35																				1
[+son,-nas]g	g	70			1			1												1		
	k	30																				1
[+nas]g	g	100			1				1												1	
	k	0																				1

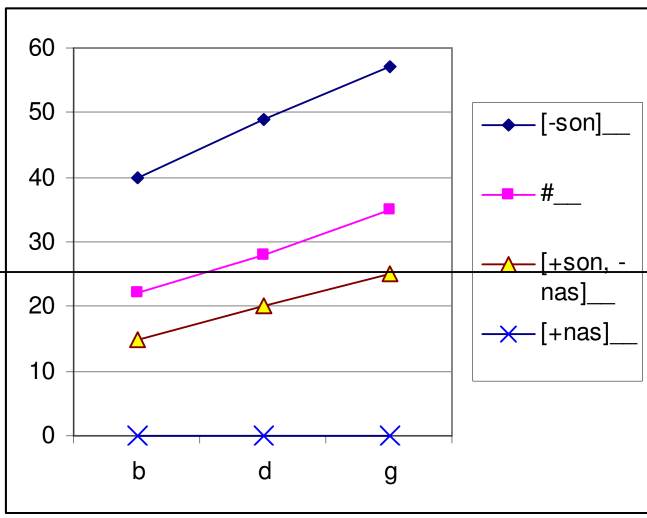
20 Results

Recall training:



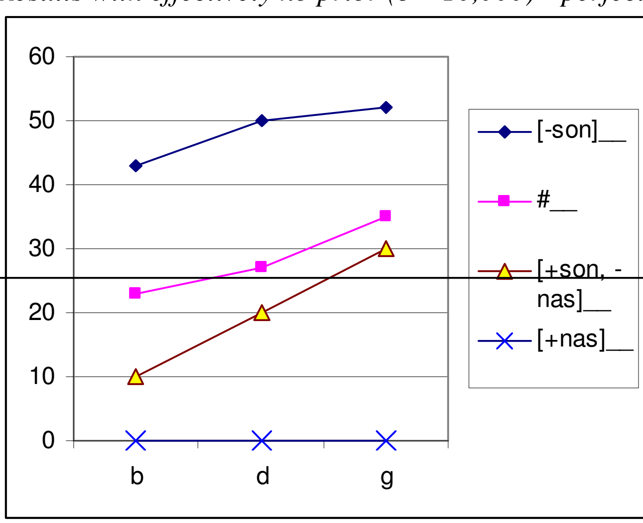
- ○ What would this picture have to look like so that any horizontal line separates some contexts from other (and doesn't distinguish place)?
- ○ So that any horizontal line separates some places from others (and doesn't distinguish context)?

Results with effectively no prior (σ^2=10,000) "row" and "column" constraints only



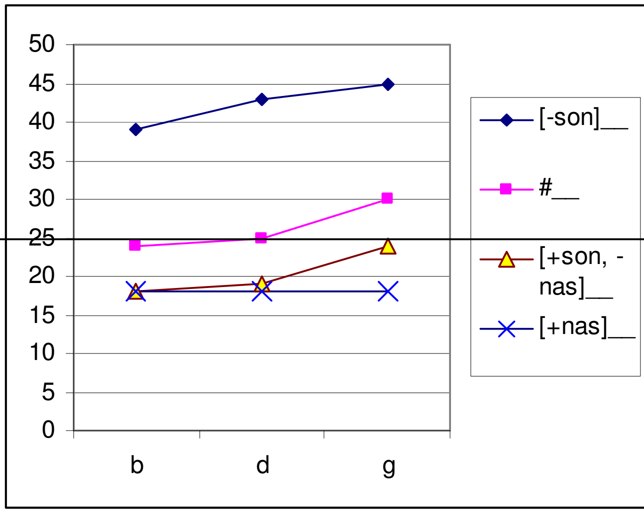
- Hayes was making the simplifying assumption that resulting grammars would have an invariable ranking.
  - ○ Discuss how this is different.

Results with effectively no prior (σ^2=10,000)—perfect learning



- There are enough constraints to fit the data perfectly.

Results with strong Gaussian prior (σ^2=0.1)



- This is looking a little better, but there are still "cell" constraints getting substantial weight.

- Unfortunately, I don't have the software to impose a square-root prior.
  - But what might we expect, in light of the row/column-constraints-only results?

21 Wrap-up

- Even with no substantive (e.g., phonetically driven) bias, a smoothing term still holds the learner back from perfectly fitting the data.
- <u>Martinian leakage</u>: If there are most context-specific and more general constraints, trends will leak from the context they start in (e.g., monomorphemes) into others (e.g., compounds)
- <u>Ryanian variationogenesis</u>: Learners exposed to non-varying data, if they don't fit too closely, will (generally) invent some variation.
  - Related case that I spared you because most of you heard it in phonology seminar last year: A learner trained on only the most "basic" stress data for Tagalog two-syllable reduplication matches the observed pattern of variation for the non-basic cases pretty well.

⇒ When we observe detailed patterns of variation, we should ask how close they are to a reasonable null hypothesis.

22 Coming up

- Formal and substantive biases on variation: articulatory ease, perceptual similarity, formal simplicity...

Reference

Lewis, M. Paul (ed.). 2009. *Ethnologue: languages of the world*. 16th ed. Dallas, TX: SIL International.

Martin, Andrew. 2007a. The evolving lexicon. University of California, Los Angeles Ph.D. Dissertation.

Martin, Andrew. 2007b. Grammars leak: how categorical phonotactics can cause gradient phonotactics. Poster presentation. Paper presented at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford University.

Martin, Andrew. 2011. Grammars leak: modeling how phonotactic generalizations interact within the grammar. *Language* 87(4). 751–770.

Ryan, Kevin M. 2010. Variable affix order: grammar and learning. *Language* 86(4). 758–791. (13 April, 2012).