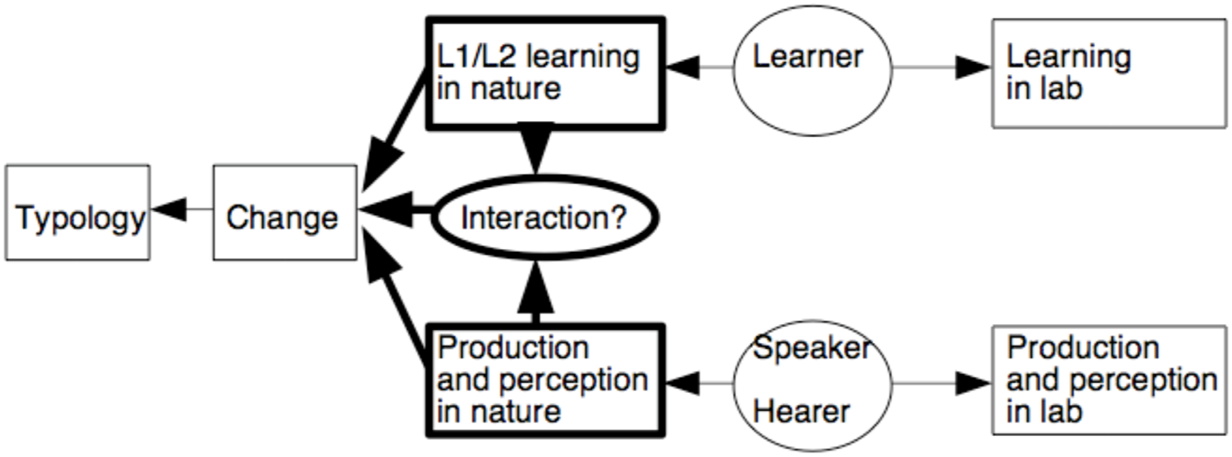
Connecting change and learning

Elliott Moreton
UNC-Chapel Hill

(1) Where we are:



(2) Main points so far:

a. Relative probabilities of typological changes completely determine long-term typological frequencies (and vice versa, up to a constant factor), if certain language-contact effects are rare enough to be safely ignored.

b. Typological change probabilities depend on analytic and channel bias in the transmission of linguistic competence between generations.

(3) What lies ahead today:

a. Relating typological change to channel and analytic bias. How quantitative dare we be?

b. Defining and quantifying channel and analytic bias. What properties must a learner have in order for them to be well-defined?

c. The logic of the typical lab-learning experiment. Assuming that it's "like" L1 or L2 learning in the relevant ways, what can analytic bias measured in the lab tell us about analytic bias in nature (and hence about relative probabilities of typological change)?

d. Concrete example: Height-height and height-voice patterns (continued from last time).

1 Model of type change

(4) Markov model (Bell, 1970; Greenberg, 1978) reduces typological bias to transition bias ("Why is $S_1 \rightarrow S_2$ more likely than $S_1 \rightarrow S_3$?"). Where does transition bias come from?

(5) Model of type change by community: Based loosely on Weinreich et al. (1968). Basic idea:

a. Speakers have one very frequent grammar, plus other options that are much rarer. Predominant grammar determines typological category.

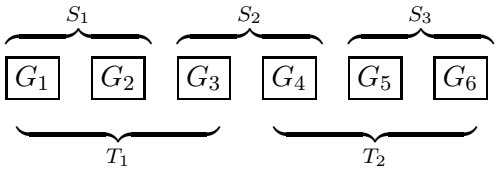
b. Change happens for social reasons only: One of the rarer grammars gets promoted to predominance for reasons which have nothing to do with its phonology.

c. Probability of promoting an optional grammar depends on the frequency of that optional grammar, which in turn depends on how likely it is to be innovated on the basis of training data produced by the incumbent predominant grammar.

d. Hence, probability of typological change ultimately depends on probability of innovating the target grammar as an optional process.

(6) Model of mature speaker:

a. There is a fixed set of available deterministic grammars $\{G_1, \ldots, G_n\}$. Different typologies classify these grammars in different ways:



b. Speakers control multiple styles, which they use depending on social context. We'll ignore all but two of these styles, the "type style" (the one that determines the typological classification of the whole language) and the "lab style" (the new style learned in the lab).

c. Variable processes: A style is completely specified by $\bar{\phi} = (\phi_1, \ldots, \phi_n)$, where ϕ_k is the probability of using G_k in that style.

G_1	G_2	G_3	G_4	G_5	G_6
S_1		S_2		S_3	

d. We can write $\phi(S_i)$ to mean the sum of all ϕ_k such that $G_k \in S_i$. A style $\bar{\phi}$ is typologically in S_i if $\phi(S_i) > \phi(S_j)$ whenever $j \neq i$ (i.e., most-probable type determines membership).

(7) Model of community of mature speakers:

a. All speakers in community have same type style $\bar{\phi}$.

b. The type style heavily favors the predominant grammar: $\phi(S_i) \gg \phi(S_j)$ for $j \neq i$. Consequently, all communities speaking a language in S_i have approximately the same type style $\bar{\phi}_i$.

c. Social promotion: An optional grammar may become predominant (= typological change) for reasons which are entirely extra-linguistic and indifferent to the content of the grammars involved. All the model can see is that

$$p_{ij} = \Pr(S_i \to S_j) = s \cdot \bar{\phi}_i(S_j)$$

where s is fixed and small, and $i \neq j$.

(8) Model of learner:

a. Learning outcome $\bar{\phi}'$ depends only on the frequencies of certain events in the training corpus $out(\bar{\phi})$.

   E.g., for GLA (Boersma, 1997; Boersma and Hayes, 2001) or Maximum Entropy (Goldwater and Johnson, 2003; Hayes and Wilson, 2008) learner, the events are violations of particular constraints (Jäger, ming); for RCD (Tesar, 1995), they are Elementary Ranking Condition rows (Prince, 2002; Magri, 2008); for parameter-setting learner, they are particular cues (Dresher, 1999).

b. $\Rightarrow$ Can represent input to the learner as $\bar{d} = (d_1, \ldots, d_n)$, and any two corpora which have the same frequencies will produce the same learning outcome. Example:

Events	CV	V	CVC	VC	CCV	...
$\bar{d}$	$d_1 = 13661$	d_2	$d_3 = 10902$	$d_4 = 160$	$d_5 = 0$	...

c. Since the corpora are so large, any two speakers in the same community will produce essentially the same $\bar{d}$ (two big samples from same frequency distribution; Law of Large Numbers). $\Rightarrow$ We can speak as if *out* were a deterministic function.

(9) Model of inter-generational transmission:

a. A given mature speaker of a language in S_i generates an output corpus with event frequencies $\bar{d}_i = out(\bar{\phi}_i)$. Ex. for final-obstruent devoicing:

Sequence:	/t./	/d./	/n./	/V./
$\bar{d}_i$:	1000	3	500	750

b. The corpus passes through the speech channel (articulation, acoustics, perception). The channel alters the frequencies in $\bar{d}_i$ by some linear function which can be expressed as a confusion matrix $\mathbf{C}$. E.g.:

Stimulus	Response			
	t.	d.	n.	V.
t.	0.99	0.01	0.00	0.00
d.	0.06	0.82	0.11	0.01
n.	0.00	0.08	0.91	0.01
V.	0.00	0.00	0.00	1.00

$$\bar{d}_i \cdot \mathbf{C} = (999, 49, 456, 753)$$

c. The learner learns from $\bar{d}_i \cdot \mathbf{C}$, i.e.,

$$\bar{\phi}'_i = lrn(out(\bar{\phi}_i) \cdot \mathbf{C})$$

   Since change happens only for social reasons (by the Weinreich et al. hypothesis), $\bar{\phi}'_i = \bar{\phi}_i$.[1]

d. Since $\bar{\phi}_i$ is dominated by S_i,

$$p_{ij} \approx s \cdot (lrn(out(S_i) \cdot \mathbf{C}))(S_j)$$

[1]This means the learner can't simply match the frequencies in the data. If it did so, then there would be runaway growth of minor processes supported by channel bias at the expense of the predominant type, leading to non-social change. The learner has to unfairly favor the predominant type to prevent this from happening (Hudson Kam and Newport, 2005; Griffiths and Kalish, 2007).

i.e., the probability of a type transition from S_i to S_j is proportional to the acquired probability of S_j, given data generated by S_i alone and transmitted via the biased channel.[2]

(10) There is no realistic hope of figuring out what s is, so we won't be able to find p_{ij}. However, we ought to be able to find the *ratio* of two transition probabilities, since the s's will cancel out. E.g.,

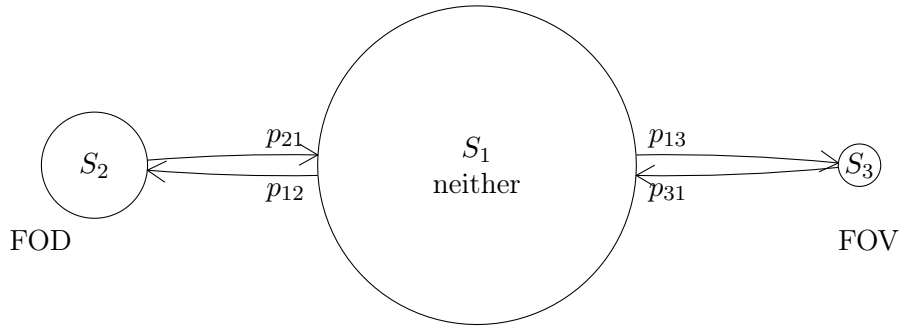
$$\frac{p_{12}}{p_{13}} = \frac{lrn(out(S_1) \cdot \mathbf{C})(S_2)}{lrn(out(S_1) \cdot \mathbf{C})(S_3)}$$

And we know from last time that what matters for typological frequency is the ratio of the transition probabilities, not their absolute size.

2 Connecting learning to channel and analytic bias

(11) Analytic bias (aka inductive bias) means different learning response to data of equal statistical quality. ⇒ To study it, we need a way to define "equal statistical quality" between two data sets.

(12) Example: Suppose we are comparing the innovation of final-obstruent devoicing and final-obstruent voicing:



Suppose the only relevant categories for the learner are those of certain syllable-final segments (ignoring alternations for simplicity):

Segment:	/t./	/d./	/n./	/a./
$\vec{d}_i \cdot \mathbf{C}$:	999	49	456	753

A learner exposed to this evidence would acquire final-obstruent devoicing (S_2) to a greater extent than final-obstruent voicing (S_3). What would data of "equal statistical quality" for final-obstruent voicing look like? Presumably:

Segment:	/t./	/d./	/n./	/a./
$\vec{d}_i \cdot \mathbf{C} \cdot \mathbf{H}_{23}$:	49	999	456	753

[2]Because of L1 effects on perception, $\mathbf{C}$ will depend to some extent on S_i, and should really be written $\mathbf{C}_i$. We'll dodge this complication today by only considering cases where it doesn't matter.

$$(999, 49, 456, 753) \cdot \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{H}_{23}} = (49, 999, 456, 753)$$

(13) Generally, we want $\mathbf{H}_{23}$ to rearrange the entries in $\bar{d}$ in such a way that

a. $\mathbf{H}_{23}$ has an inverse, so we can swap cues going in the other direction.

b. $\mathbf{H}_{23}$ preserves the statistical representativeness of $\bar{d}$ with respect to the types:

(i) $\Pr(\bar{d} = out(S_1)) = \Pr(\bar{d} \cdot \mathbf{H}_{23} = out(S_1))$

(ii) $\Pr(\bar{d} = out(S_2)) = \Pr(\bar{d} \cdot \mathbf{H}_{23} = out(S_3))$

(14) It follows that

$$\begin{aligned} \frac{p_{12}}{p_{13}} &= \frac{lrn(out(S_1) \cdot \mathbf{C})(S_2)}{lrn(out(S_1) \cdot \mathbf{C})(S_3)} \\ &= \frac{lrn(out(S_1) \quad \cdot \mathbf{C})(S_2)}{lrn(out(S_1) \cdot \mathbf{H}_{23} \cdot \mathbf{C})(S_3)} \\ &= \underbrace{\frac{lrn(out(S_1) \cdot \mathbf{C} \quad)(S_2)}{lrn(out(S_1) \cdot \mathbf{C} \cdot \mathbf{H}_{23})(S_3)}}_{A_{23}} \cdot \underbrace{\frac{lrn(out(S_1) \quad \cdot \mathbf{C} \cdot \mathbf{H}_{23})(S_3)}{lrn(out(S_1) \cdot \mathbf{H}_{23} \cdot \mathbf{C} \quad)(S_3)}}_{C_{23}} \end{aligned}$$

a. A_{23} compares learning response to data of equal statistical quality as a function of the learning target. If you feed the learner with S_2-like and S_3-like data of the same statistical quality, does it learn the respective patterns equally well?

b. C_{23} expresses difference in learning the same target as a function of the difference in how the channel treats final [t] vs. final [d] .

c. Both kinds of bias are quantified according to their effect on *learning*. Channel bias isn't directly measurable from spectrograms or confusion experiments (though it may be indirectly measurable from them).

(15) This seems to give us much of what we want:

a. Neat separation between effects of analytic and channel bias on relative probabilities of diachronic change

b. Common unit for measuring and comparing analytic and channel bias

c. Suggestions for experiments to measure them.

(16) However: We can also rewrite p_{12}/p_{13} as

$$\frac{p_{12}}{p_{13}} = \underbrace{\frac{lrn(out(S_1) \cdot \mathbf{C} \cdot \mathbf{H}_{23}^{-1})(S_2)}{lrn(out(S_1) \cdot \mathbf{C} \quad)(S_3)}}_{A_{32}} \cdot \underbrace{\frac{lrn(out(S_1) \cdot \mathbf{H}_{23}^{-1} \cdot \mathbf{C} \quad)(S_2)}{lrn(out(S_1) \quad \cdot \mathbf{C} \cdot \mathbf{H}_{23}^{-1})(S_2)}}_{C_{32}}$$

If the channel- and analytic-bias terms of (14) equal their correspondents, then it makes sense to speak of "the" respective contributions of channel and analytic bias to favoring $S_1 \rightarrow S_2$ transitions relative to $S_1 \rightarrow S_3$ transitions. Otherwise, we have a presupposition failure.

(17) Under what circumstances can we speak of "the" analytic and channel biases? The channel biases will be equal if the analytic ones are, so we need

$$\underbrace{\frac{lrn(out(S_1)\cdot\mathbf{C}\qquad)(S_2)}{lrn(out(S_1)\cdot\mathbf{C}\cdot\mathbf{H}_{23})(S_3)}}_{A_{23}} = \underbrace{\frac{lrn(out(S_1)\cdot\mathbf{C}\cdot\mathbf{H}_{23}^{-1})(S_2)}{lrn(out(S_1)\cdot\mathbf{C}\qquad)(S_2)}}_{A_{32}}$$

(18) $\Rightarrow$ We want the disparity in learning response to two **H**-isomorphic sets of training data to be the same if the two data sets are not too different from each other (differ by about as much as the channel effect).

(19) We could get that to happen by deleting the "if" clause, so that we require of the learner that

$$\frac{lrn(\bar{d}\qquad)(S_2)}{lrn(\bar{d}\cdot\mathbf{H}_{23})(S_3)} = \frac{lrn(\bar{d}'\qquad)(S_2)}{lrn(\bar{d}'\cdot\mathbf{H}_{23})(S_3)}$$

for all $\bar{d}, \bar{d}'$.

(20) Won't work:

a. If $\bar{d}$ is $out(S_2)$, then $\bar{d}\cdot\mathbf{H}_{23}$ is $out(S_3)$, and $lrn(\bar{d})(S_2) = 1 = lrn(\bar{d}\cdot\mathbf{H}_{23})$, so the ratio is 1.

b. But then the ratio has to be 1 for all other $\bar{d}'$, and the learner can't show any analytic bias at all!

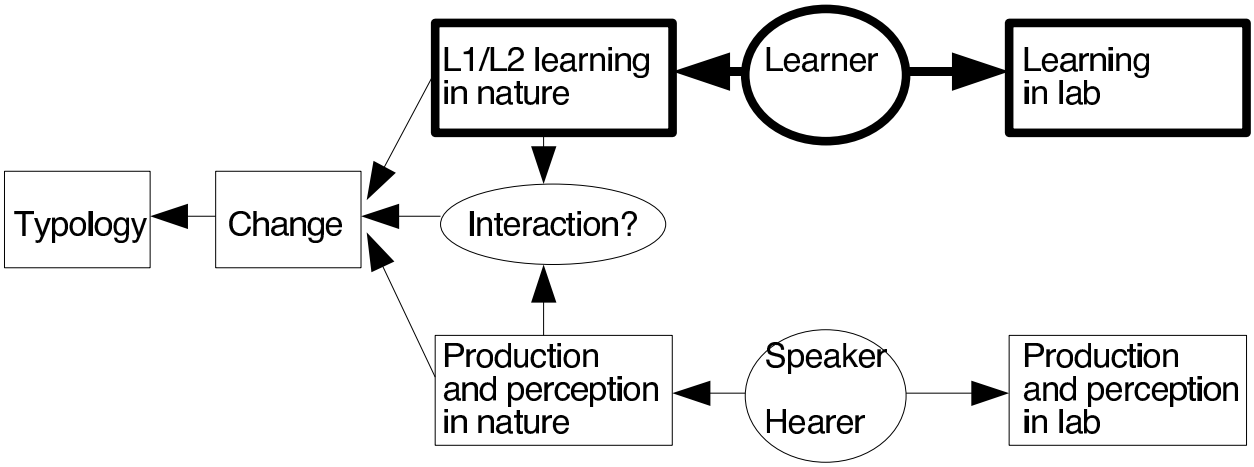
(21) Solution: Better to use $\text{odds}(p) = p/(1-p)$, which for small p is $\approx p$, and which can preserve a constant ratio across the full range of probabilities from 0 to 1:

$$\frac{\text{odds}(lrn(\bar{d}\qquad)(S_2))}{\text{odds}(lrn(\bar{d}\cdot\mathbf{H}_{23})(S_3))} = \frac{\text{odds}(lrn(\bar{d}'\qquad)(S_2))}{\text{odds}(lrn(\bar{d}'\cdot\mathbf{H}_{23})(S_3))}$$

for all $\bar{d}, \bar{d}'$.

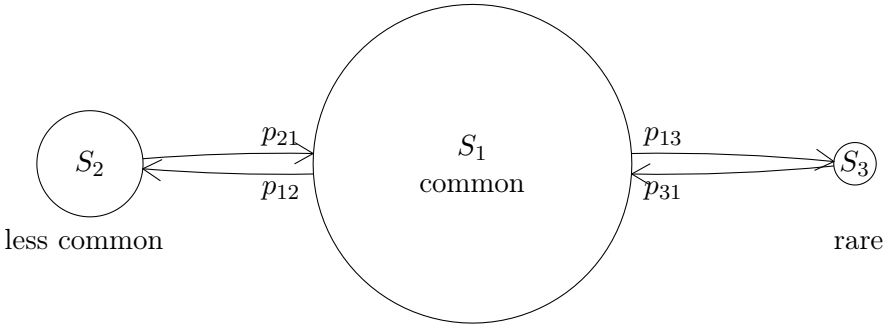
3 Connecting natural learning to laboratory learning

(22) Where we are:



(23) Typical lab experiment to measure analytic bias:

a. Participant shows up having already been trained on L1 corpus $\bar{d}_1$ belonging to type S_1, and the types being compared are S_2 and S_3. The implicit typology is usually the "Martian typology" (big planet, two little moons):



b. Familiarize on one of two data sets, $\bar{\delta}_2$ (consistent with S_2 only) or $\bar{\delta}_3$ (consistent with S_3 only), where $\bar{\delta}_3 = \bar{\delta}_2 \cdot \mathbf{H}$.

c. The lab situation creates a new social context with an associated style. Participants may weight lab experience more heavily than L1 experience, so the learning outcome is

$$\bar{\phi}_{lab,i} = lrn(r \cdot \bar{d}_1 + (1 - r) \cdot \bar{\delta}_i)$$

d. Test using some task where Pr(correct) is

(i) p_t when the participant applies whichever of S_2 or S_3 they were trained on. Usually experimenters try for $p_t = 1$.

(ii) p_g when they guess (apply S_1, use a non-linguistic strategy, etc.). In a two-alternative forced-choice test, for instance, the aim is to have $p_g = 1/2$.

(24) If the learner satisfies (21), then it is straightforward to get from the responses to estimates of *the* analytic bias:

a. Let $l_i = \bar{\phi}_{lab,i}(S_i)$, i.e., l_i is how much credence participants put in S_i after begin trained in the lab on $\bar{\delta}_i$.

b. Let p_i be Pr (choose correct response when trained on δ_i). Then

$$\begin{aligned} p_i =& l_i \cdot p_t + (1 - l_i) \cdot p_g \\ =& l_i \cdot (p_t - p_g) + p_g \end{aligned}$$

c. The experiment gives us an empirical estimate $\hat{p}_i$ for p_i, from which we can calculate an estimate $\hat{l}_i$ for $lrn(\bar{d}_1 + \bar{\delta}_i)(S_i, \text{exp})$:

$$\hat{l}_i = \begin{cases} \dfrac{\hat{p}_i - p_g}{p_t - p_g} & \text{if } \hat{p}_i > p_g \\ 0 & \text{otherwise} \end{cases}$$

and so

$$\text{odds}(\hat{l}_i) = \begin{cases} \dfrac{\hat{p}_i - p_g}{p_t - \hat{p}_i} & \text{if } \hat{p}_i > p_g \\ 0 & \text{otherwise} \end{cases}$$

d. The estimated analytic bias is thus

$$\hat{A}_{23} = \frac{\text{odds}(\hat{l}_2)}{\text{odds}(\hat{l}_3)}$$

(25) Connecting that back to to typological change, we predict that

$$\begin{aligned} \frac{p_{12}}{p_{13}} =& \hat{A}_{23} \cdot \hat{C}_{23} \\ =& \frac{\text{odds}(\hat{l}_2)}{\text{odds}(\hat{l}_3)} \cdot \hat{C}_{23} \end{aligned}$$

(26) Backup plan, in case learner really doesn't satisfy (21), and analytic bias varies depending on the training data: We can at least reasonably hope that the *direction* of the bias never changes, and that $\forall \bar{d}, \bar{d}'$,

$$lrn(\bar{d})(S_2) > lrn(\bar{d} \cdot \mathbf{H})(S_3) \iff lrn(\bar{d}')(S_2) > lrn(\bar{d}' \cdot \mathbf{H})(S_3)$$

and hence that

$$A_{23} > 1 \iff \text{odds}(\hat{l}_2) > \text{odds}(\hat{l}_3)$$

---

4 Example: height-height and height-voice interactions (cont'd)

(27) Patterns (from previous handout):

HH pattern Predictive dependency between vowel height in adjacent syllables (harmony or disharmony in height between V_1 and V_2 in $V_1C_0V_2$). Seems to be rather frequent.

HV pattern Predictive dependency between vowel height and "voicing" (= phonetic voicing, aspiration, or fortis/lenis contrast) of immediately-following obstruent (V_1C_1). Seems to be rare.

(28) Assumed typology (some states ignored; see previous handout):

(29) Phonological survey results, with 95% CIs:

	$\frac{\pi_2}{\pi_2+\pi_3}$			Equvalent $\frac{\pi_2}{\pi_3}$		
Sample	Lower	Mean	Upper	Lower	Mean	Upper
Strict (7:0)	0.768	0.938	1	3.31	15.1	∞
Lax (14:2)	0.656	0.853	0.973	1.91	5.80	36.0

(30) From last time:

$$\frac{\pi_2}{\pi_3} = \frac{p_{12}}{p_{13}} \cdot \frac{p_{21}}{p_{31}}$$

implies

$$\frac{p_{12}}{p_{13}} \text{ or } \frac{p_{21}}{p_{31}} \geq \sqrt{\frac{\pi_2}{\pi_3}}$$

$\Rightarrow$ We hypothesize that $p_{12}/p_{13} \geq \sqrt{\pi_2/\pi_3}$.

I.e., in order for HH to outnumber HV by as much as it does in real life, one of two things has to be true: Either HH is more likely to be innovated from S_1, or, once innovated, it is less likely to be lost, by a certain minimum margin in either case. We are testing the hypothesis that it is more likely to be innovated (by at least the margin).

(Why are we focusing on innovations rather than extinctions?)

(31) Linking hypothesis, from (25):

$$\frac{p_{12}}{p_{13}} = A_{23} \cdot C_{23}$$

so

$$\sqrt{\pi_2/\pi_3} < \frac{\text{odds}(l_2)}{\text{odds}(l_3)} \cdot C_{23}$$

But what is C_{23}?

(32) Phonetic-precursor surveys (see Moreton (2008) for details and Yu (ress) for critique):

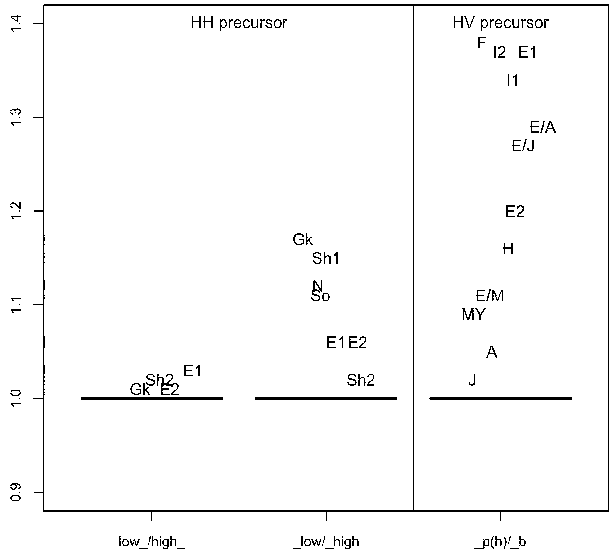
a. Find studies where vowel F_1 is measured in the relevant contexts.

b. Identify contexts likeliest to raise or lower target F_1

c. Effect of context is defined to be (F_1 in raising context)/(F_1 in lowering context).

d. If F_1 was measured at multiple points, the one closest to the context was used.

(33) HH and HV precursors:[3]



(34) $\Rightarrow$ *Production* effect on F_1 is larger in HV than HH precursor. If we *assume* that perceptual confusions follow production differences, then we can infer that

$$\begin{aligned}\Rightarrow C_{23} =& \frac{lrn(out(S_1) \quad \cdot \mathbf{C} \cdot \mathbf{H}_{23})(S_3)}{lrn(out(S_1) \cdot \mathbf{H}_{23} \cdot \mathbf{C} \quad)(S_3)} \\ <& 1\end{aligned}$$

(35) So our hypothesis becomes

$$\begin{aligned}\sqrt{\pi_2/\pi_3} <& \frac{\text{odds}(l_2)}{\text{odds}(l_3)} \cdot (\text{something} < 1) \\ <& \frac{\text{odds}(l_2)}{\text{odds}(l_3)}\end{aligned}$$

[3] A = Arabic; E1, E2 = English; E/A, E/J, E/M = L2 English (L1 = Arabic, Japanese, Mandarin); F = French; Gk = Greek; H = Hindi; I1, I2 = Italian; J = Japanese; MY = Mòbà Yoruba; N = Ndebele; Sh1, Sh2 = Shona; So = Sotho.

I.e., if HH is more likely (by a given margin) to be innovated, then it must enjoy an advantage in either analytic or channel bias. The phonetic results indicate that the advantage is not in channel bias, so it must be in analytic bias, and hence analytic bias must favor HH by the given margin.

(36) Stimuli: MBROLA-synthesized $C_1V_1C_2V_2$ words with inventory /t k d ɡ/ /i u æ ɔ/. Two patterns:

a. "HH pattern": Vowels agree in height, instantiating a height-harmony pattern.

b. "HV pattern": V_1 high iff C_2 voiced, instantiating what *would* be a phonologization of the HV precursor.

(37) The **H** isomorphism relating the HH and HV stimuli simply swaps the value of the voicing feature of C_2 with that of the height feature of V_2: /taku/ → /taɡɔ/ , etc. Thus, the number of /ak/ stimuli in an HV set equals the number of /a…ɔ/ stimuli in the corresponding HH set.

(38) Experimental paradigm (based on Moreton (2008, Exps. 1 and 2)):

a. *Study Phase*: Listen to pattern-conforming words through headphones, repeat into microphone. 32 words × 4 repetitions, randomized in blocks.

Pattern conformity		Training condition	
HH	HV	HH	HV
+	+	16	16
+	−	16	
−	+		16
−	−		

b. *Test Phase*: Listen to pairs of new words, choose the one that you think is "a word of the language you studied". 32 pairs in two counterbalanced blocks of 16, random orders in block and pair. Each pair pits one pattern-conforming item against one pattern-nonconforming item:

Pattern conformity					Studied pattern	
HH	HV		HH	HV	HH	HV
+	+	*vs.*	−	−	16	16
+	−	*vs.*	−	+	16	16

(39) Properties of this design:

a. For half of the Test pairs, the correct response depends on the Study pattern; for the other half, it does not. Allows effects of learning to be separated from those of pre-existing preferences.

b. Does *not* test generalization to new vowels or new combinations of vowels (i.e., does not distinguish between learning vowel harmony and learning a list of vowel-vowel sequences). (Same applies to speakers of real vowel-harmony languages too, of course.)

(40) Participants: 18 native speakers of American English. None had studied or otherwise learned a language with vowel harmony. One explicitly noticed pattern (post-experiment questionnaire) and was replaced.

(41) Theory applies in ideal circumstances, but real-life experiment is not ideal. Use logistic regression to model out nuisance factors (mixed-effects mode with Participant as a random effect).

This exp. was part of a series of 6 that were very similar to each other. Regression coefficients used in this analysis were those that could not be eliminated from *at least one* of the 6 analyses.

Coefficient	Estimate	SE	z	$Pr(>\mid z \mid)$	
(Intercept)	0.27419	0.19609	1.39830	0.162024	
Studied HH	0.71606	0.27884	2.56804	0.010228	*
$V_1 = V_2$	–0.25962	0.20536	–1.26420	0.206160	
2nd half	–0.27877	0.24170	–1.15339	0.248750	
Studied HH × *2nd half*	–0.05977	0.35390	–0.16889	0.865882	
HH-nonconforming	0.10146	0.13140	0.77217	0.440015	
1st in pair	0.46502	0.17679	2.63042	0.008528	**

(42) We want a confidence interval (95%, let's say) for $A_{23} = \frac{\text{odds}(l_2)}{\text{odds}(l_3)}$.

a. Two coefficients matter: $\beta_0 = 0.27419$ (s.e. $= 0.19609$) and $\beta_1 = 0.71606$ (s.e. $= 0.27884$). Randomly generated 1,000,000 of each from the respective normal distributions.

b. $p_{HV} = p_3 = \exp(\beta_0)/(1 + \exp(\beta_0))$; $p_{HH} = p_2 = \exp(\beta_0 + \beta_1)/(1 + \exp(\beta_0 + \beta_1))$.

c. odds(l_2), odds(l_3), and A calculated as in (24)

d. About 8.1% of the A_{23}s were infinite (happens when $\beta_0 < 0$, and hence $l_3 = 0$).

e. Median was 5.66; (geometric) mean of finite values was 5.90. 95% of values were above 2.08.

(43) Comparison with hypothesis:

Quantity		Lower	Mean	Upper
$\sqrt{\frac{\pi_2}{\pi_3}}$	strict	1.81	3.89	∞
	lax	1.34	2.41	6.00
$\frac{\text{odds}(l_2)}{\text{odds}(l_3)}$		1.44	5.66	∞

⇒ Our backup position is strongly supported; there is analytic bias favoring S_{HH} over S_{HV}.

At a more precise quantitative level, the results are numerically encouraging ($\hat{A}_{23}$ is larger than the other two means), but the confidence intervals are much too wide to tell for sure which one is bigger.

5 Discussion

(44) What can we conclude from the example? To what extent have we *explained* the prevalence of HH over HV patterns in real-world phonologies?

(45) Can we run the predictive chain in the other direction?

(46) What kind of learner would satisfy the requirement in (21)? Is it a reasonable learning model for phonology? Do these results support the hypothesis that real learners are like that?

(47) Suppose your preferred learning model *doesn't* satisfy (21). How would you derive predictions from it about lab and natural biases? How precise could those predictions be?

(48) Channel bias, like analytic bias, is stated in terms of its effect on the learner (see 14 above). Can channel bias, so defined, be measured in the lab?

(49) Are we guaranteed to be able to find an $\mathbf{H}_{23}$ that satisfies the requirements in (13)? Under what circumstances might we not? What implications would that have—could we still, for example, compare the analytic or channel bias between S_2 and S_3?

References

Bell, A. (1970). *A state-process approach to syllabicity and syllabic structure*. Ph. D. thesis, Stanford University.

Boersma, P. (1997). Functional Optimality Theory. *Proceedings of the Institute of Phonetic Sciences of University of Amsterdam 21*, 37–42.

Boersma, P. and B. Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry 32*, 45–86.

Dresher, B. E. (1999). Charting the learning path: cues to parameter setting. *Linguistic Inquiry 30*(1), 27–67.

Goldwater, S. J. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Erkisson, and O. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pp. 111–120.

Greenberg, J. H. (1978). Diachrony, synchrony, and language universals. In J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (Eds.), *Universals of human language, volume 1, method and theory*, pp. 61–91. Stanford, California: Stanford University Press.

Griffiths, T. L. and M. L. Kalish (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science 31*(3), 441–480.

Hayes, B. and C. Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*(3), 379–440.

Hudson Kam, C. L. and E. Newport (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development 1*, 151–195.

Jäger, G. (forthcoming). Maximum Entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (Eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan*. Stanford: CSLI.

Magri, G. (2008). Linear methods in Optimality Theory: a convergent incremental algorithm that performs both promotion and demotion. MS, Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology 25*(1), 83–127.

Prince, A. (2002). Entailed ranking arguments. MS, Rutgers Optimality Archive (ROA–500).

Tesar, B. (1995). *Computational Optimality Theory*. Ph. D. thesis, University of Colorado.

Weinreich, U., W. Labov, and M. Herzog (1968). Empirical foundations for a theory of language change. In W. P. Lehmann and Y. Malkiel (Eds.), *Directions for historical linguistics: a symposium*, pp. 97–195. Austin: University of Texas Press.

Yu, A. C. L. (in press). Tonal effects on perceived vowel duration. MS, University of Chicago. To appear in: Papers in Laboratory Phonology 10.