Linguistic Institute
Boulder, Colo.

Formally biased phonology: Complexity in learning and typology[1]

2011, July 13

Elliott Moreton
University of North Carolina

Joe Pater
University of Massachusetts

1 Introduction

(1) Featurally simple phonological patterns:

a. Are learned more easily than more complex patterns (LaRiviere et al. 1974, 1977, and others cited below; see Moreton & Pater, submitted, for a review).

b. Are over-represented typologically relative to more complex patterns (Clements 2003; Moreton 2008; see also Mielke 2004, Ch. 6).

(2) The aims of *formally biased phonology*:

a. To empirically investigate the role of formal complexity in learning and typology

b. To explicitly formalize simplicity biases in learning and their impact on typology

Some others who have advocated or explored some part of this research agenda: Bach and Harms (1972); Hale and Reiss (2000); Hayes (1999), as well as Gary Dell's talk this morning.

(3) An example of simplicity bias in learning - Saffran and Thiessen's (2003) phonotactic learning/word segmentation study:

a. *Training Phase 1.* 9-month-olds exposed to isolated words of shape $C_1VC_2.C_1VC_2$, where C_1 and C_2 are each limited to a set of three consonants.

b. *Training Phase 2.* Exposed to 4 new words in a continuous stream, with only two fitting the pattern from Phase 1.

c. *Test Phase.* Listening time measured for each of the words from Training Phase 2, presented in isolation.

d. *Results 1.* When the pattern was *featurally simple*, with voiceless [p, t, k] in one position, and voiced [b, d, g] in the other, the infants displayed a *novelty preference* for non-conforming items.

e. *Results 2.* When the pattern was *featurally complex*, with [p, d, k] in one position, and [b, t, g] in the other, there was *no significant difference* in listening time between conforming and non-conforming items.

[1]The authors gratefully acknowledge the advice and suggestions already contributed by Anne Pycha, Jen Smith, Colin Wilson, and the participants in Linguistics 751 at the University of Massachusetts in Spring 2011, especially John Kingston, John McCarthy, Claire Moore-Cantwell, Presley Pizzo, and Robert Staubs. Thanks also to Andrew Cohen for drawing our attention to the Configural Cue Model. Work supported in part by NSF Grant # 0813829.

(4) An example of simplicity bias in typology - statistical dependence between [b] and [g] in the phoneme inventories of UPSID-92 (John Kingston p.c., based on Maddieson and Precoda (1992); see further Clements (2003) on feature economy):

a.

	/b/	no /b/
/g/	244	11
no /g/	43	153

b. $\chi^2 = 260, d.f. = 1, p < 0.01$

c. Languages tend to have either both [b] and [g] or neither: it is especially unlikely for a language to have [g] without [b].

d. More generally, inventories tend to avoid "holes" and "bumps": A segment is more likely if all of its feature values are shared by other segments. In other words, languages tend to avoid proliferating features; they tend toward feature economy (Martinet, 1968; Clements, 2003)

(5) These and similar data from learning and typology have yet to be given a fully explicit account.

a. Chomsky and Halle's (1968: 334) evaluation procedure prefers rules that use fewer features, but it was originally proposed only as a means of choosing between two analyses of a single set of data. It is possible that a learning algorithm incorporating the evaluation procedure or other Minimum Description Length principle could be used to account for simplicity biases in learning and typology, but this remains to be shown.

b. Complex patterns can be learned, just with more difficulty, and are only typologically underrepresented, not absent. These sorts of probabilistic tendencies fall out of the scope of most phonological theories.

(6) Today's talk:

a. A demonstration that a simplicity bias in learning *emerges* when the Perceptron Update Rule (Rosenblatt, 1962, 110, Mitchell, 1997, Ch 4.4) is used with an appropriately structured constraint set.

b. An examination of the effects of featural complexity on phonological learning, and the relation to non-linguistic pattern learning (e.g., Shepard et al., 1961)

c. A model of emergent phonological complexity bias: an implementation of the Configural Cue Model (Gluck and Bower 1988a) as Perceptron learning of Maximum Entropy phonotactics (Hayes and Wilson, 2008).

d. A demonstration that learning biases can have an effect on typology when incomplete learning is perpetuated and amplified through agent-based interaction.

2 Emergent simplicity bias

(7) To illustrate the emergence of a simplicity bias, we here adopt Hayes and Wilson's (2008) Maximum Entropy phonotactic model of grammar, which defines a probability distribution over the space of possible word forms. A simple example:

a.

	[+Vce] −4	[-Vce] 4	[+Vce]∧[+Cor] 8	
[b]	1			$H = -4, p < .001$
[d]	1		1	$H = 4, p = .25$
[g]	1			$H = -4, p < .001$
[p]		1		$H = 4, p = .25$
[t]		1		$H = 4, p = .25$
[k]		1		$H = 4, p = .25$

b. Three positive constraints [+Vce], [-Vce], and [+Vce]∧[+Cor] assign a reward of 1 to consonants that are voiced, voiceless, and voiced and coronal, respectively.

c. The constraints' weights are given underneath the constraint names.

d. The value labeled H at the end of each row provides a numerical score of each consonants' well-formedness, or *Harmony*, the weighted sum of rewards.

e. The p value for each consonant is the probability assigned by the grammar, which is proportional to the exponential of the consonant's Harmony.

f. In this toy example, the linguistic universe is the set of six consonants, and the probabilities are those that a hypothetical language assigns to each of them.

(8) The Perceptron update procedure for MaxEnt phonotactics:

a. A single learning datum is sampled from the target distribution (observed).

b. A single datum is sampled from the distribution defined by the learner's grammar (expected).

c. The difference is taken between the vectors of constraint scores (observed - expected), and the resultant difference vector is scaled by the learning rate.

d. The scaled difference vector is added to the vector of constraint weights to get the updated values.

(9) Example languages from the [p,t,k,b,d,g] universe, with constraint sets. The three consonants in each language have equal probability in the learning data. Constraints reward each single feature and two-feature conjunction present in the language (constraint [t] abbreviates [+Vce]∧[+Cor]):

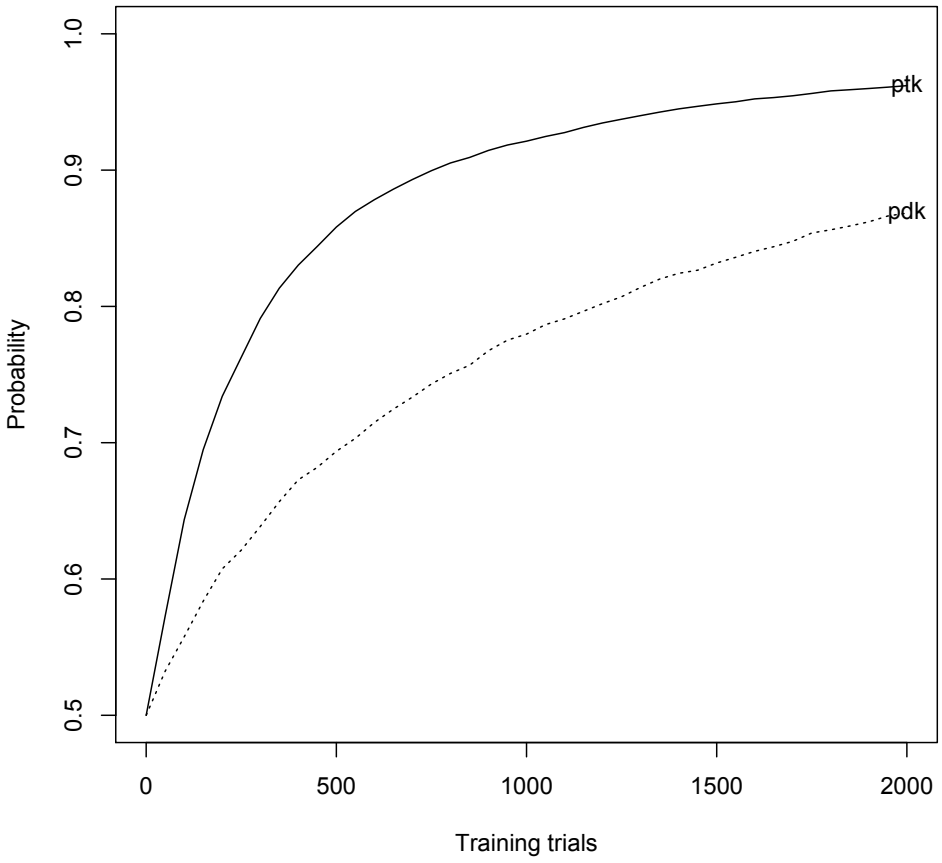
a. *Language 1 Data.* [p, t, k]

b. *Language 1 Constraints.* [-Vce], [+Lab], [+Cor], [+Dor], [p], [t], [k]

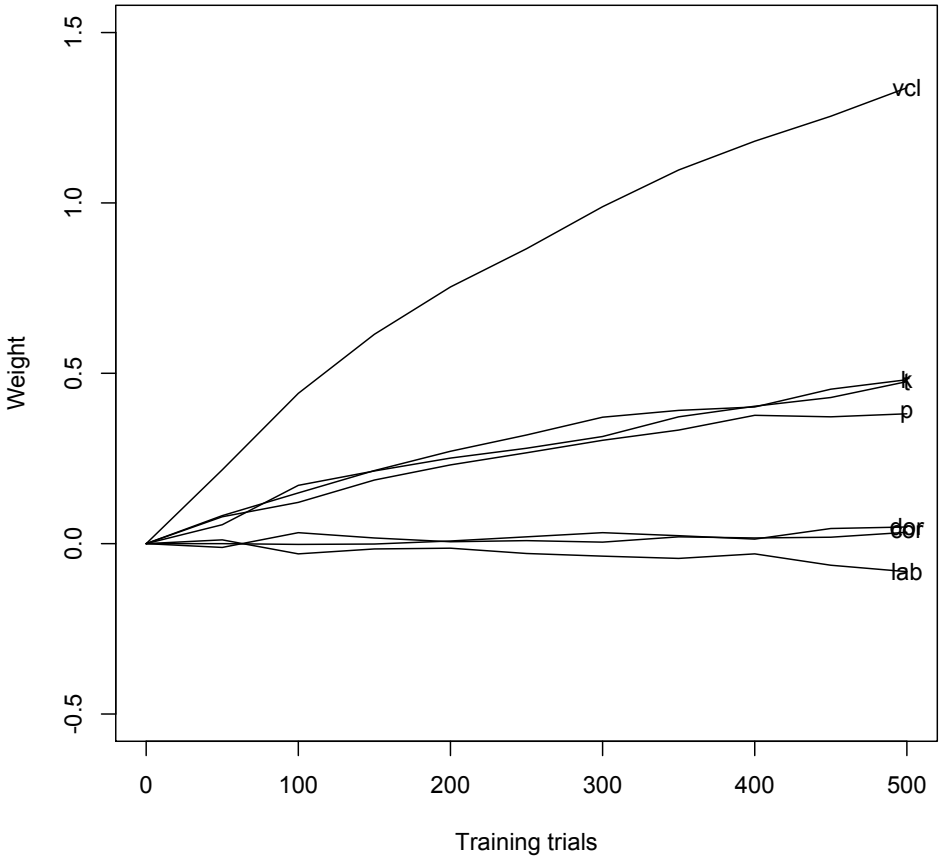
c. *Language 2 Data.* [p, d, k]

d. *Language 2 Constraints.* [-Vce], [+Vce], [+Lab], [+Cor], [+Dor], [p], [t], [k]

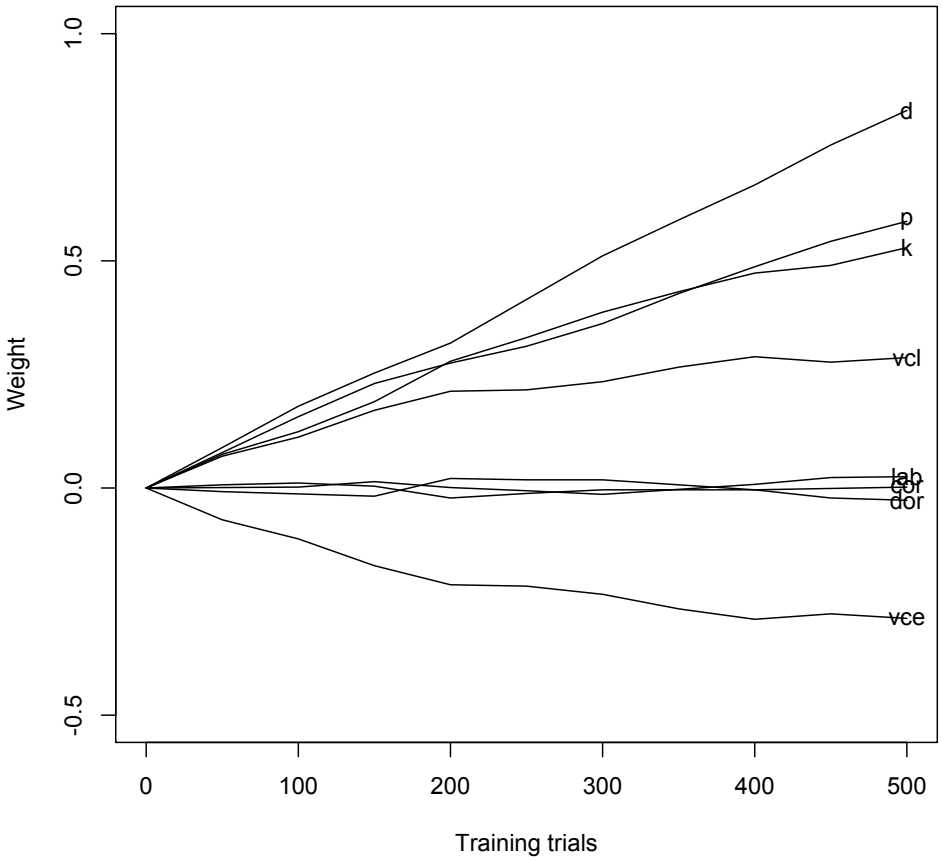
(10) Constraint weights started at zero, and the learning rate was set at 0.01. This graph shows the probability assigned to the observed forms over the course of 2000 learning trials, at 50 trial intervals. At every point in learning, the [p,t,k] learner assigns higher probability to the observed forms in its language than the [p,d,k] learner does.



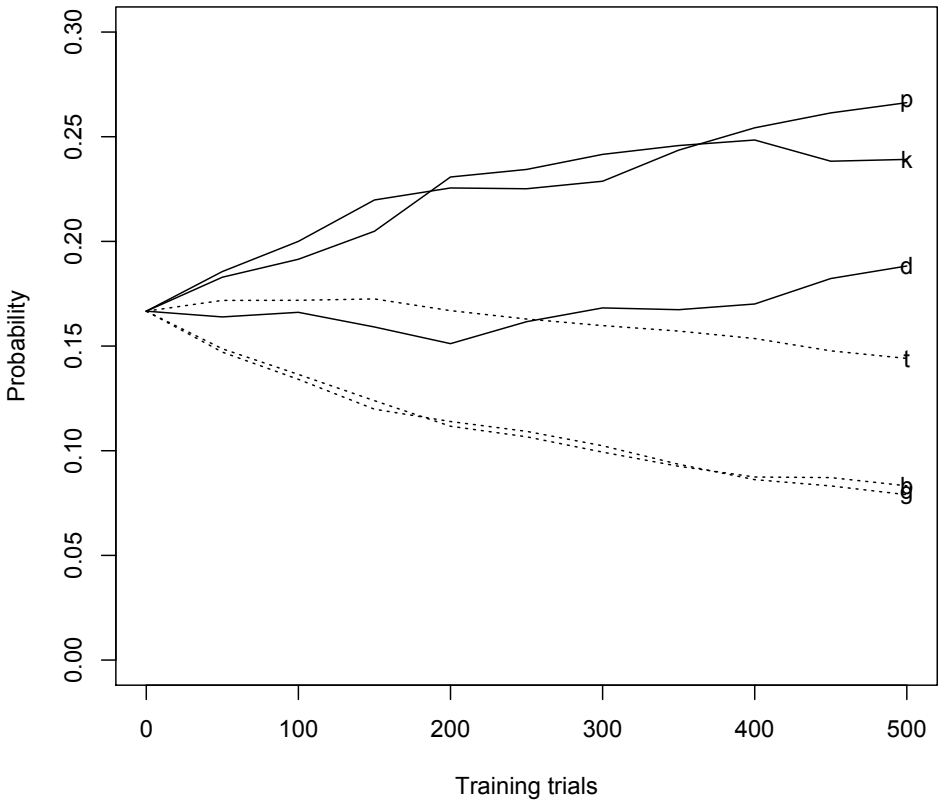
(11) The next three graphs show why [p,t,k] is easier to learn than [p,d,k] under these assumptions. First, for every mismatch between observed and expected, the [p,t,k] learner adds positive weight to [-Vce]. It thus rises quickly, as shown in this graph of constraint weights over the first 500 trials.



(12) There is no single constraint that picks out the observed forms for the [p, d, k] learner. Because voiceless consonants predominate, [-Vce] does get a positive weight, and [+Vce] a negative weight, whose joint effects must be overcome by a high weighted [d] constraint.



(13) This leads to some *over-generalization* in early stages of learning (as in Rumelhart and McClelland (1986)): compare [t] to [d] in trials 50-200. We'll come back to over-generalization in the typology modeling:



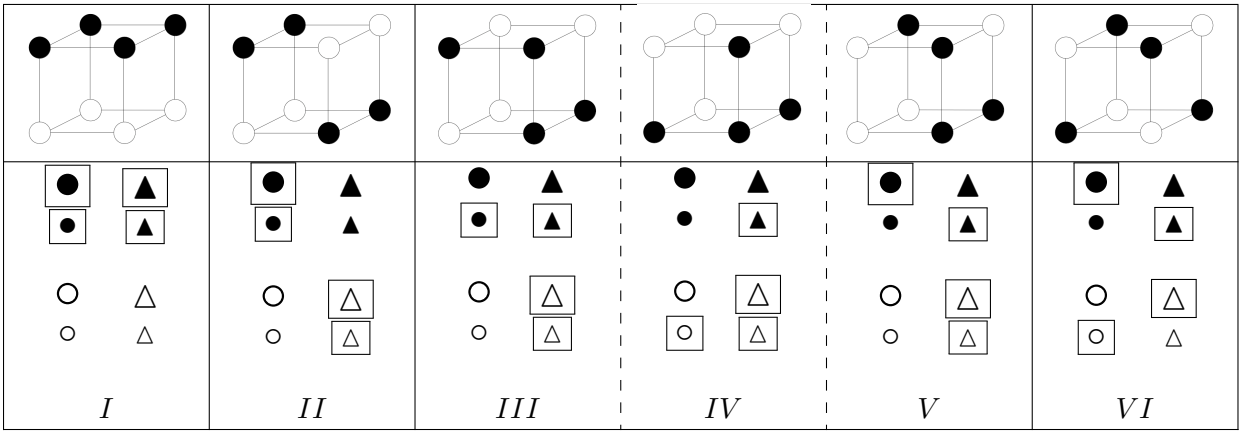
(14) The main result is an *emergent simplicity bias*: the system that can be described with a single voicing feature is learned faster than the one that requires both place and voicing. This bias is hard-wired only in the sense that it depends on the structure of the constraint set:

a. It requires the general single feature constraint [–Vce]: an alternative model that contains only the two-feature constraints learns both languages at the same rate.

b. The assumption that constraint sets include such general constraints is uncontroversial, but not without content: it implies that learning involves abstraction from the featural make-up of individual forms. For example, there is no observed form that is [–Vce], without also being specified for some place feature.

3 Formal complexity and empirical difficulty

(15) A stimulus space described by three binary-valued features can be divided into two equal-sized categories in only six ways (ignoring reflections and rotations) (Shepard et al., 1961):



Examples are shown for patterns defined over the three features color (black vs. white), shape (circle vs. triangle), and size (large vs. small).

a. Type I: Only color matters.

b. Type II: Only color and shape matter.

c. Types III–V: All three features, matter, but some subsets can be described with fewer (e.g., white triangles).

d. Type VI: Every subset requires all three features.

(16) Non-linguistic pattern-learning studies:

a. Supervised learning: See a stimulus, guess whether it's positive or negative, get feedback (i.e., the correct answer).

b. Routinely give the same results:

$$\text{I} < \text{II} < \{\text{III, IV, V}\} < \text{VI}$$

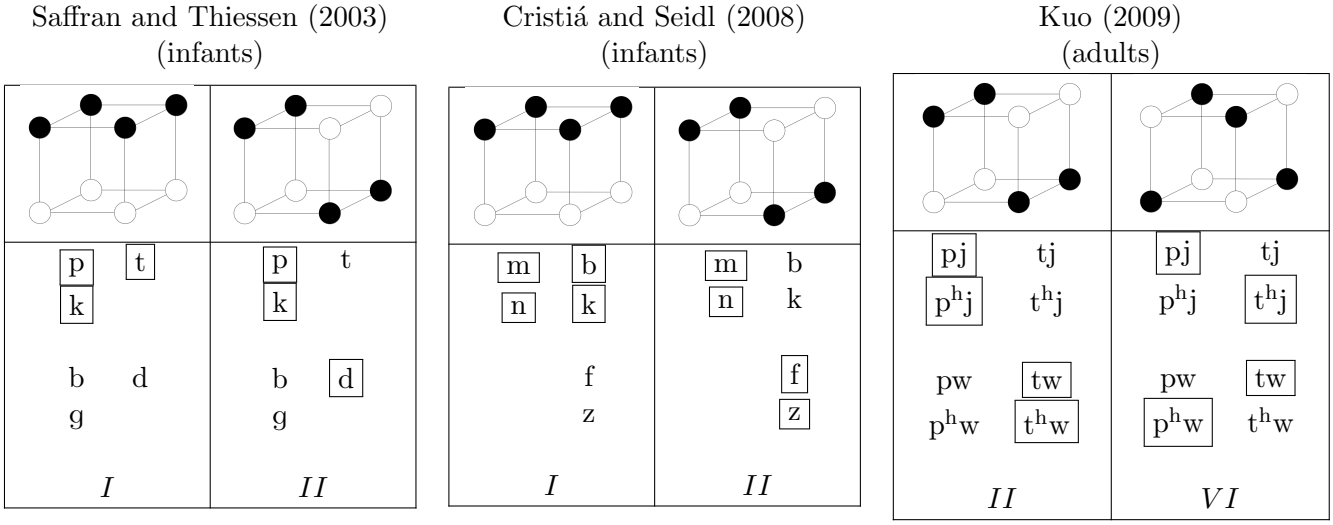
(Shepard et al., 1961; Neisser and Weene, 1962; Nosofsky et al., 1994; Feldman, 2000; Love, 2002; Smith et al., 2004).

(17) Learning artificial phonological patterns by ear: Usually unsupervised familiarize-and-test paradigm.

a. Familiarization: Hear positive stimuli only

b. Test: Distinguish positive from negative stimuli

(18) Have tested the I < II < VI subset and gotten the same results. (See also Pycha et al. (2003), not shown here, for II < VI in adults.)



(19) The I < II < VI results are reassuring, but not too surprising — it is hard to devise a plausible learning model that would predict any other order. What about the other classic finding, II < {III, IV, V} < VI? The other types haven't yet been tested in phonotactic learning.

(20) Experiment with all six types (Moreton & Pertsova, in prep.).

a. Stimuli: MBROLA-synthesized $C_1V_1C_2V_2$ words with inventory /t k d g/ /i u æ ɔ/ (Moreton, 2008):

Feature	Stimulus segment			
	C_1	V_1	C_2	V_2
voiced	±		±	
Coronal	±		±	
high		±		±
back		±		±

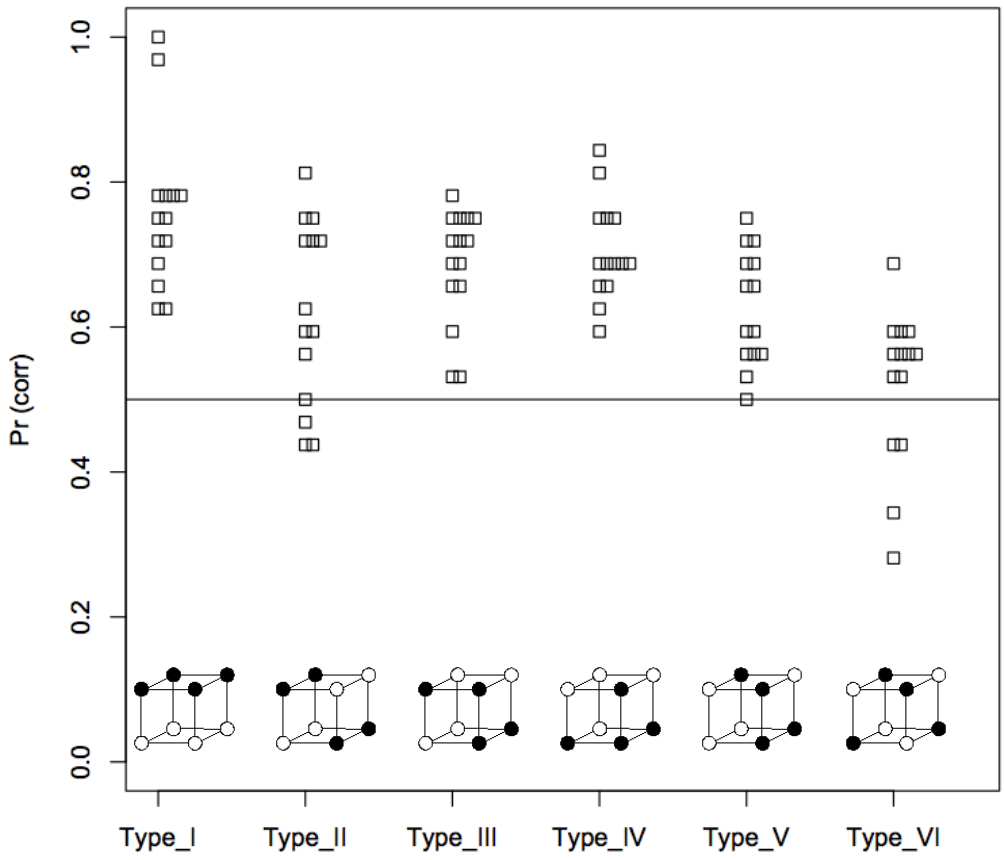
Consonants				Vowels			
k	t	g	d	æ	ɔ	i	u
−	−	+	+				
−	+	−	+				
				−	−	+	+
				−	+	−	+

b. Each participant randomly assigned to one of Types I–VI.

c. For each participant, randomly choose 3 of the 8 stimulus features then randomly map those onto the 3 logical features defining the problem type to yield the artificial language. (Pattern is almost sure to lack a natural-language analogue.)

d. Participants were told they would learn to pronounce words in an artificial language, and then be tested on ability to recognize words in that language. They listened to and repeated aloud 32 positive stimuli 4 times over, then heard 32 pairs of new stimuli (one positive, one negative) and tried to identify the positive one.

(21) Interim results (14 participants per condition), shown as proportion correct responses in the test phase:



(22) Analysis by logistic regression, with Type II as reference category.

Type	Estimate	se	p value		Notes
II (ref)	0.49182	0.09736	< 0.0001	***	II better than chance
I	0.65499	0.14724	< 0.0001	***	I easier than II
III	0.28696	0.13835	0.03807	*	III easier than II
IV	0.38112	0.14220	0.00736	**	IV easier than II
V	0.02854	0.13794	0.83609		V similar to II
VI	-0.41142	0.13573	0.00244	**	VI harder than II

a. **Replicates** I < II < VI order found previously with non-linguistic stimuli and with phonological stimuli.

b. **Reverses** the usual non-linguistic finding of II < {III, IV}. The pattern that used only two features turned out to be *more* difficult than the ones that used three.

(23) **Interim summary**: Featural complexity seems to affect both phonological and non-linguistic pattern learning, but in different ways, suggesting different learning mechanisms.

(24) *Caveat*: Obviously, what counts as "complex" depends on what descriptive primitives you have available. We're using the term in order to maintain continuity with previous linguistic and psychological work. "Complexity bias" can be mentally replaced with "formal structural bias" if preferred (Hale and Reiss, 2000).

---

4 Modelling the learning bias

(25) Need to explain

a. How the learning bias comes about. Is it hard-wired? Emergent?

b. Why it is different from the usual non-linguistic findings.

(26) Explanations that won't work:

a. Any model built to predict the classic I < II < { III, IV, V } < VI order (e.g., Kruschke, 1992; Love et al., 2004).

b. *Boolean complexity* (Feldman, 2000, 2004, 2006), which predicts I < {II, III} < IV < V < VI (Lafond et al., 2007).

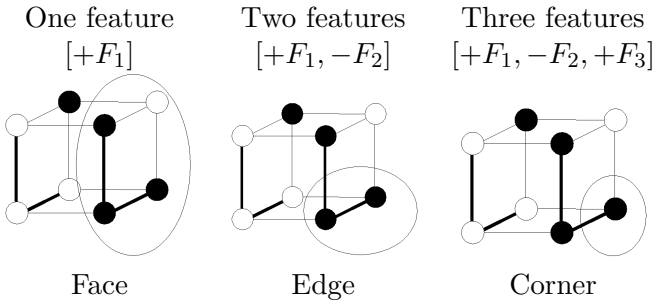
c. *Linear separability*, which Love (2002) has proposed as an explanation for why II can be harder than IV in unsupervised non-linguistic learning. This wouldn't explain why III is also harder than II.

(27) Instead, we will reach back to an emergent model which was proposed for non-linguistic category learning, and was rejected because it overestimated the difficulty of Type II: the Configural Cue Model of Gluck and Bower (1988b,a).

4.1 The Configural Cue Model (Gluck and Bower, 1988b,a)

(28) Originally implemented as a single-layer feedforward neural net. The key ideas were

a. Unbiased conjunctive constraints: There is a constraint for every possible conjunction of feature values over the whole set:



b. Error-driven learning: On each training trial, adjust the influence of a constraint up or down in proportion to its effect on the output error ("Delta Rule", "Widrow-Hoff Rule").

(29) Straightforward adaptation to current OT-derived phonological learning models (Pater et al., 2008)

a. Unbiased conjunctive constraints: One for every conjunction of +, −, or "don't care" feature values over the whole stimulus set ($[-F_2], [+F_2 - F_3], [+F_1 + F_2 - F_3], \ldots$]).

b. Error-driven learning: There are incremental learning algorithms analogous to the Delta Rule for Stochastic OT (Boersma, 1997; Boersma and Hayes, 2001; Magri, 2008), Harmonic Grammar (Pater, 2008; Boersma and Pater, 2008), and Maximum Entropy grammar (Jäger, 2007).

(30) The present simulation used a Maximum Entropy learner updated using the Perceptron Update Rule, implemented in Praat (Boersma and Weenink, 2011)

a. Constraints are "positive", i.e., they favor the outputs they describe. They can apply penalties if given negative weights.

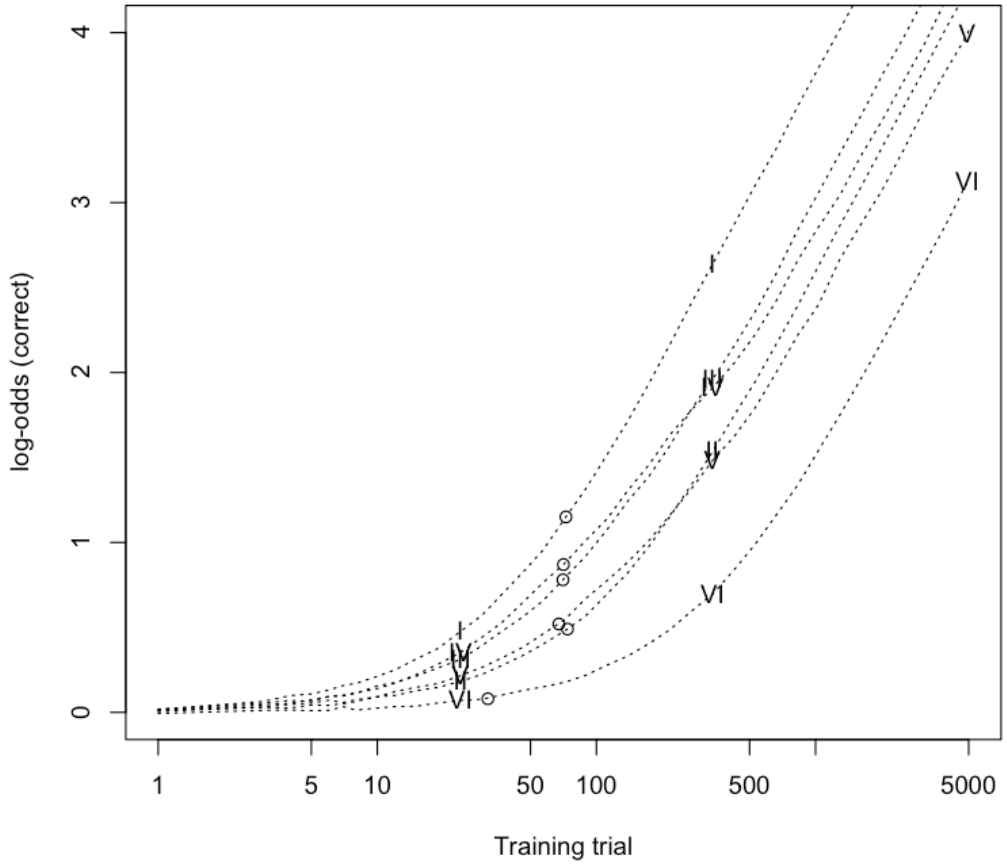
b. Input is always /X/; output candidate set is all eight possible stimuli. Training distributions (desired output distributions) for each Shepard problem type:

	Stimulus (output candidate)							
Type	− − −	− − +	− + −	− + +	+ − −	+ − +	+ + −	+ + +
I					0.25	0.25	0.25	0.25
II	0.25	0.25					0.25	0.25
III			0.25	0.25		0.25		0.25
IV				0.25		0.25	0.25	0.25
V	0.25			0.25		0.25		0.25
VI	0.25			0.25		0.25	0.25	

c. 15 replications of each Type. 5000 training trials; 20 "chews" per trial; plasticity 1/2000. All constraints initially had weight 0.

d. Humans gave 2AFC responses, while model divided probability mass among the 8 stimuli. So: For each possible test pair (S_+, S_-), define model's predicted probability of choosing S_+ as $\Pr(S_+)/(\Pr(S_+) + \Pr(S_-))$ (Luce, 1959, 20 ff.).

(31) Log-odds[2] of proportion correct, as a function of training duration:



(32) Comparison with human data from Figure (21):

a. Human data was collected after a fixed amount of training. Simulation was tested throughout training.

b. Qualitative human results (I < {III, IV} < {II, V} < VI) are consistent with the model throughout.

c. Circles in Figure (31) show when the model best matches the human data quantitatively.

d. Humans are behaving the way the model does after about 60 training trials, except in the case of Type VI.

[2]Log-odds odds is related to proportion p by $l = \ln(p/(1-p))$. When $p = 0.5, l = 0$, and $l \to \infty$ as $p \to 1$.

4.2 The importance of partially valid cues

(33) Why does the model behave this way? Gluck and Bower (1988a, 188–189) point to the importance of "partially valid cues", i.e., constraints which

a. Apply only to a proper subset of the stimulus set, and

b. Are always valid about stimuli in that subset

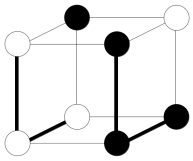
A constraint can be valid by favoring positive stimuli, or by disfavoring negative ones.

(34) Because learning is error-driven, constraints can gain influence (larger positive or negative weights) only on trials where they contribute to the decision. For Types II–VI,

a. Each single-feature constraint (cube face) is *relevant* on half of trials, but is *valid* on at most 75% of those.

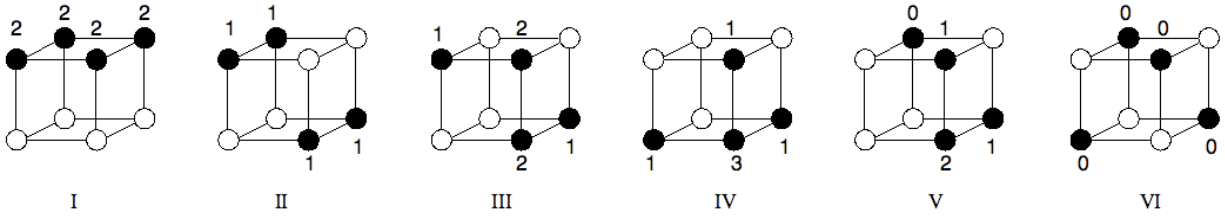
b. Each three-feature constraint (cube corner) is *relevant* on just 1/8 of trials, though it is always *valid* when it is relevant.

c. Some two-feature constraints (cube edges) are *relevant* on 1/4 of trials, and *valid* on all of those. These constraints describe cube edges that connect two positive or two negative stimuli. We will call these "valid edges" for short. Type V has four valid edges, shown in bold:



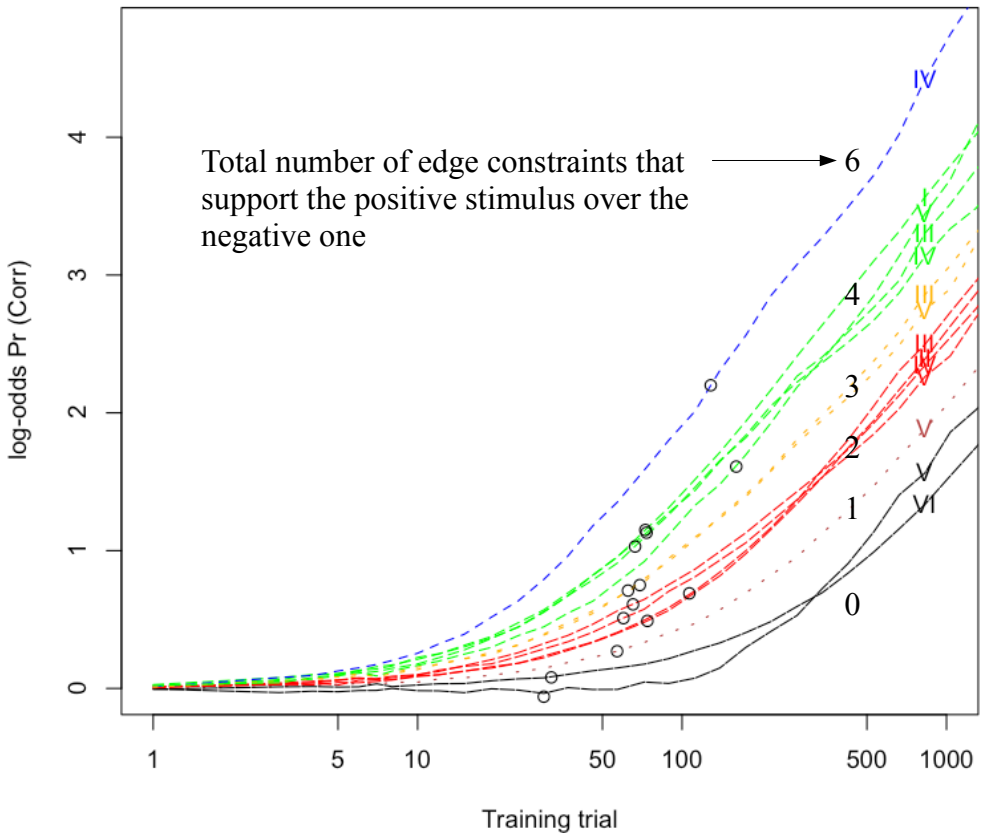
⇒ Valid edge constraints should gain influence fastest and hence should dominate responses.

(35) Within each concept, the individual stimuli are supported by different numbers of valid edges. This illustration shows only how many support each positive stimulus; the negative stimuli are symmetrical.



(36) In the simulation, stimuli supported by the same number of valid edges are learned (gain or lose probability mass) at about the same rate, *regardless of which concept they occur in.*

These curves show log-odds of a correct two-alternative forced-choice test response at each training stage, broken down by the number of valid edges favoring the positive stimulus over the negative one:



⇒ Responses are dominated by the valid edges, as suggested by Gluck and Bower (1988a). The other constraints get promoted slower and exert less influence:

a. The single-feature constraints $[\alpha F_1]$ (cube faces) aren't valid (except in Type I).

b. The three-feature constraints $[\alpha F_1, \beta F_2, \gamma F_3]$ (cube corners) are all valid, but they are relevant only half as often as the valid edges.

(37) Comparison with the human data from Moreton & Pertsova (in prep.):

a. Log-odds of a correct human response for each test pair, grouped within each type according to how many valid edges favored the correct response:

Edge	Type					
Constraints	I	II	III	IV	V	VI
6				2.20		
4	1.15		1.03	1.13	1.61	
3			0.71		0.75	
2		0.49	0.68	0.61	0.51	
1					0.27	
0					-0.06	0.08

⇒ Difficulty of a test pair depends more on the edge constraints than on the pattern type.

b. Plotted as circles in Figure (36). Again, humans act approximately like the model trained for about 60 trials.

(38) **Interim summary**:

a. The human phonotactic-learning results are inconsistent with most models of non-linguistic category learning.

b. They are fairly close to the predictions of a model which has been rejected for non-linguistic pattern learning, Gluck and Bower (1988a)'s Configural Cue Model.

c. In that model, the complexity biases biases emerge from two facts:

(i) Error-driven learning favors fast promotion of constraints which characterize coherent subsets of the pattern (constraints which are relevant to, and valid on, subsets described by conjunctions of few features)

(ii) Less-complex patterns afford more such coherent subsets (more partially-valid cues).

5 How learning can skew typology

(39) Here we provide a simple demonstration of how the simplicity bias we have been exploring in language learning could have typological consequences.

(40) (Martinet, 1968) suggests a learning-based explanation of feature economy:

a. "...each of [the features] being more frequent in speech, speakers will have more occasions to perceive and produce them, and they will get anchored sooner in the speech of children. A phoneme that is integrated into one of those bundles of proportional oppositions which we call 'correlations' will in principle be more stable than a non-integrated phoneme...."

b. A learning bias does not *automatically* lead to a typological skew. Insofar as difficult systems are only learned more slowly, they will still be learned, given enough time.

(41) Learning biases can have an impact on language change, and hence typology, if the results of incomplete learning are transmitted from one learner to another

a. Effect of a simplicity bias on language change perhaps first demonstrated by Hare and Elman (1995) in their connectionist model of historical change in English verb inflection.

b. See Zuraw (2003) and Wedel (2011) for overviews of agent-based modeling in phonology

c. See Griffiths et al. (2008) for a Bayesian model of iterated learning applied to simplicity biases in concept learning and for related experimental results.

(42) Our agent-based modeling set-up:

a. Two learners are each exposed to the same target distribution for some number of trials (childhood).

b. Some number of interaction trials follow in which one learner is randomly chosen, who produces an output for a randomly chosen input/word, from which the other learner learns (adolescence).

(43) Learning again uses the Perceptron update procedure, and the grammar model again uses MaxEnt probability distributions, but now over OT-style input-output candidate sets, as in Goldwater and Johnson (2003):

a.

/P/	[+Asp] 4	[+Vce]∧[+Lab] 2	[+Vce]∧[+Cor] 6	
[b]		1		$H = 2, p = .12$
[p^h]	1			$H = 4, p = .88$

b.

/T/	[+Asp] 4	[+Vce]∧[+Lab] 2	[+Vce]∧[+Cor] 6	
[d]			1	$H = 6, p = .88$
[t^h]	1			$H = 4, p = .12$

(44) For each place/input/word, a choice is made between voicing and aspiration. Our learners got one of two categorical distributions in childhood, and a universal constraint set (phonetic symbols again abbreviate two-feature conjunctions; candidate sets have fixed place so no single-feature place constraints are included):

a. *Language 1 Data.* /P/ → [b], /T/ → [d], /K/ → [g]

b. *Language 2 Data.* /P/ → [p^h], /T/ → [t^h], /K/ → [g]

c. *Constraints.* [+Asp] , [+Vce], [p], [t], [k], [p^h], [t^h], [k^h]

(45) Simulation parameters:

a. Starting weight zero, zero minimum maintained through learning

b. Learning rate 0.01 for single-feature constraints, 0.001 for two-feature constraints

c. 1000 learning trials in childhood, 100,000 in adolescence

(46) Results in terms of number of sibling pairs (out of ten) that assign [g] $p > .7$, and [k^h] $p > .7$:

a. *Language 1.* [g]: 8 [k^h]: 0

b. *Language 2.* [g]: 4 [k^h]: 3

(47) With other constraint sets that have the same abstract specific-general structure, this account generalizes to other instances of "systemic simplicity", such as stress regularization, and consistent headedness across types of syntactic phrase.

---

6 Conclusions

(48) Compared to analogous proposals about phonetic naturalness, the hypothesis of an inductive bias regarding formal complexity has not received detailed elaboration or testing. Current state of knowledge:

a. *Quantitative typology*: How does formal complexity affect cross-linguistic pattern frequency? (Clements, 2003; Mielke, 2004; Moreton, 2008)

b. *Lab learning*: How does complexity affect learnability in the lab? In fact, the evidence for complexity effects is much more robust than that for phonetic-naturalness effects (several studies, reviewed in Moreton & Pater, submitted).

c. *Implementation*: What (explicit) learning algorithms *make* a complexity bias happen? (Hayes, 1999; Hayes and Wilson, 2008).

d. *Modelling*: How well do the learning algorithms account for the lab data? For the typological data?

e. *Cognition*: How are the foregoing related to the vast psychological literature on complexity in non-phonological pattern learning?

(49) Main new points:

a. Formal complexity impedes unsupervised phonological learning differently from supervised non-linguistic learning. What matters is less the number of relevant features than the parsability of the pattern into coherent subsets.

b. The biases in the human phonological learning data are surprisingly well matched by a model which uses the Perceptron rule and has a constraint for every such possible coherent subset (Gluck and Bower, 1988a).

c. These biases can lead to imperfect learning, causing a more-complex pattern to be mis-learned as a simpler one in an iterated-learning simulation.

References

Bach, E. and R. T. Harms (1972). How do languages get crazy rules? In R. P. Stockwell and R. K. S. Macaulay (Eds.), *Linguistic change and generative theory*, Chapter 1, pp. 1–21. Bloomington: Indiana University Press.

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.

Boersma, P. and J. Pater (2008, May). Convergence properties of a gradual learning algoritm for Harmonic Grammar. MS.

Boersma, P. and D. Weenink (2011). PRAAT Version 5.2.08. Software, www.praat.org.

Chomsky, N. and M. A. Halle (1968). *The sound pattern of English*. Cambridge, Massachusetts: MIT Press.

Clements, G. N. (2003). Feature economy in sound systems. *Phonology 20*(3), 287–333.

Cristiá, A. and A. Seidl (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development 4*(3), 203–227.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature 407*, 630–633.

Feldman, J. (2004). How surprising is a simple pattern? quantifying "eureka!". *Cognition 93*(3), 199–224.

Feldman, J. (2006). An algebra of human concept learning. *Journal of mathematical psychology 50*, 339–368.

Gluck, M. A. and G. H. Bower (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language 27*, 166–195.

Gluck, M. A. and G. H. Bower (1988b). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General 117*, 227–247.

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.

Griffiths, T. L., M. L. Kalish, and S. Lewandowsky (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B (Biological Sciences) 363*(1509), 3503–3514.

Hale, M. and C. A. Reiss (2000). 'Substance abuse' and 'dysfunctionalism': current trends in phonology. *Linguistic Inquiry 31*(1), 157–169.

Hare, M. and J. L. Elman (1995, July). Learning and morphological change. *Cognition 56*(1), 61–98.

Hayes, B. (1999). Phonetically driven phonology: the role of optimality in inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, and K. Wheatly (Eds.), *Functionalism and Formalism in Linguistics*, Volume 1: General Papers, pp. 243–285. Amsterdam: John Benjamins.

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

Jäger, G. (2007). 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*, pp. 467–479. Stanford, California: CSLI Publications.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review 99*, 22–44.

Kuo, L. (2009). The role of natural class features in the acquisition of phonotactic regularities. *Journal of psycholinguistic research 38*(2), 129–150.

Lafond, D., Y. Lacouture, and G. Mineau (2007). Complexity minimization in rule-based category learning: revising the catalog of Boolean concepts and evidence for non-minimal rules. *Journal of Mathematical Psychology 51*, 57–75.

LaRiviere, C., H. Winitz, J. Reeds, and E. Herriman (1974). The conceptual reality of selected distinctive features. *Journal of Speech and Hearing Research 17*(1), 122–133.

LaRiviere, C., H. Winitz, J. Reeds, and E. Herriman (1977). Erratum: The conceptual reality of selected distinctive features. *Journal of Speech and Hearing Research 20*(4), 817.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review 9*(4), 829–835.

Love, B. C., D. L. Medin, and T. M. Gureckis (2004). SUSTAIN: a network model of category learning. *Psychological Review 111*(2), 309–332.

Luce, R. D. (2005 [1959]). *Individual choice behavior: a theoretical analysis*. New York: Dover.

Maddieson, I. and K. Precoda (1992). Syllable structure and phonetic models. *Phonology 9*, 45–60.

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.

Martinet, A. (1968). Phonetics and linguistic evolution. In B. Malmberg (Ed.), *Manual of phonetics*, pp. 464–487. North-Holland.

Mielke, J. (2004). *The emergence of distinctive features*. Ph. D. thesis, Ohio State University.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill International Editions.

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

Neisser, U. and P. Weene (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology 64*(6), 640–645.

Nosofsky, R. M., M. A. Gluck, T. J. Palmeri, S. C. McKinley, and P. Gauthier (1994). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition 22*(3), 352–369.

Pater, J. (2008). Gradual learning and convergence. *Linguistic Inquiry 39*(2), 334–345.

Pater, J., E. Moreton, and M. Becker (2008, November). Simplicity biases in structured statistical learning. Poster presented at the Boston University Conference on Language Development.

Pycha, A., P. Nowak, E. Shin, and R. Shosted (2003). Phonological rule-learing and its implications for a theory of vowel harmony. In M. Tsujimura and G. Garding (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)*, pp. 101–114.

Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington, D.C.: Spartan Books.

Rumelhart, D. and J. McClelland (1986). On learning the past tenses of English verbs. In J. McClelland and D. Rumelhart (Eds.), *Parallel Distributed Processing*, Volume II, pp. 216–271. MIT Press.

Saffran, J. R. and E. D. Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology 39*(3), 484–494.

Shepard, R. N., C. L. Hovland, and H. M. Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs 75*(13, Whole No. 517).

Smith, J. D., J. P. Minda, and D. A. Washburn (2004). Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General 133*(3), 398–404.

Wedel, A. (2011). Self-organization in phonology. In E. A. H. Marc van Oostendorp, Colin J. Ewen and K. Rice (Eds.), *The Blackwell Companion to Phonology*, pp. 130–147. Blackwell.

Zuraw, K. (2003). Probability in language change. In S. J. Rens Bod, Jennifer Hay (Ed.), *Probabilistic Linguistics*, pp. 139–176. Cambridge, MA: MIT Press.