Workshop on Grammar Induction

May 14, 2010
Cornell University

Constraint induction and simplicity bias in phonotactic learning

Elliott Moreton[1]
University of North Carolina, Chapel Hill

1 Introduction

(1) Context:

a. Language typology is determined by the relative rates at which phonological patterns are innovated and extinguished (Bell, 1971; Greenberg, 1978), which in turn depend on interaction between the learner's inductive biases and the variation in the learning data (Myers, 2002; Moreton, 2008a).

b. ⇒ To understand typology, we need to find out what those biases are, and how they affect the acquisition of phonological patterns.

(2) This talk is about modelling two inductive biases related to the featural complexity of patterns. Preview of main points:

a. Learning a phonological pattern is easier when only a single feature is needed to characterize it, whether within or between stimuli. What the features actually are doesn't seem to matter. (Becker et al., 2007; Moreton, 2008a). Models of human phonotactic learning need to capture these facts.

b. Non-phonological learning shows a similar advantage, at least for between-stimulus complexity.

c. All these effects can linked via the Delta Rule, an incremental error-driven learning rule which figures in models of both phonological (Boersma, 1997; Boersma and Hayes, 2001; Pater, 2008; Magri, 2008; Boersma and Pater, 2008) and non-phonological learning (Rescorla and Wagner, 1972; Gluck and Bower, 1988a,b; Kruschke, 1992; Love et al., 2004).

d. The indifference of the simplicity bias to featural content means it has to apply to induced constraints. A working model in the Harmonic Grammar framework is presented in which constraints are induced by an evolutionary algorithm, with reproductive fitness being determined by the Delta Rule. The model exhibits the within- and between-stimulus simplicity biases.

2 Simplicity bias in phonotactic learning

(3) *Paradigmatic (between-stimuli) simplicity bias*: A phonological category is often easier to learn (and never harder) when it is definable in terms of a single phonetic feature, or a small number of them, than when its phonetic categorization is complex or arbitrary.

[1]I owe thanks to many people for discussion of the ideas and facts presented here, especially Adam Albright, Joe Pater, Jen Smith, Anne-Michelle Tessier, and Alan Yu, as well as audiences at UMass/Amherst, the Workshop on Computational Modelling of Sound Pattern Acquisition at the University of Alberta, and MIT. The modelling work is part of a collaboration with Joe Pater and Michael Becker. Email may be addressed to `moreton@unc.edu`.

E.g., LaRiviere et al. (1974); Healy and Levitt (1980); Pycha et al. (2003); Saffran and Thiessen (2003); Peperkamp et al. (2006); Pycha et al. (2007); Cristiá and Seidl (2008); Kuo (2009)

(4) Example: LaRiviere et al. (1974). Participants heard *Ca* syllables and learned to sort them into two categories by immediate feedback. Performance depended on how the categories were constructed:

Separable			Arbitrary	
[+strid]	[–strid]		—	—
[ʃɑ]	[ðɑ]		[ʃɑ]	[ðɑ]
[dʒɑ]	[lɑ]		[lɑ]	[dʒa]
[zɑ]	[kɑ]		[zɑ]	[kɑ]
[sɑ]	[hɑ]		[hɑ]	[sɑ]
$p_{corr} = 0.79$		$>$	$p_{corr} = 0.55$	

(5) *Syntagmatic (within-stimulus) simplicity bias*: Phonological patterns involving within-stimulus dependencies are easier if the dependency is between two instances of the same feature.

Examples: Wilson (2003); Moreton (2008a). Not found by Seidl and Buckley (2005, Exp. 2).

(6) Example from own unpublished data. Stimuli: MBROLA-synthesized $C_1V_1C_2V_2$ words with inventory /t k d g/ /i u æ ɔ/. Five patterns:

Condition	C_1	V_1	C_2	V_2
HV (baseline)		[α high]	[α voiced]	
HH		[α high]		[α high]
VV	[α voiced]		[α voiced]	
HB		[α high]		[α back]
PV	"[α velar]"		[α voiced]	

Two conditions in each experiment: HV as baseline, and one of the others for comparison. Participants (9 per condition) listen to and repeat pattern-conforming words, then listen to pairs of novel words (one pattern-conforming, one not) and choose the one you think is "a word of the language you studied".

(7) Results (logistic-regression coefficients, 0 = no effect on odds of choosing pattern-conforming test word): HH and VV (single-feature dependency) better than HB and PV (two-feature dependency).

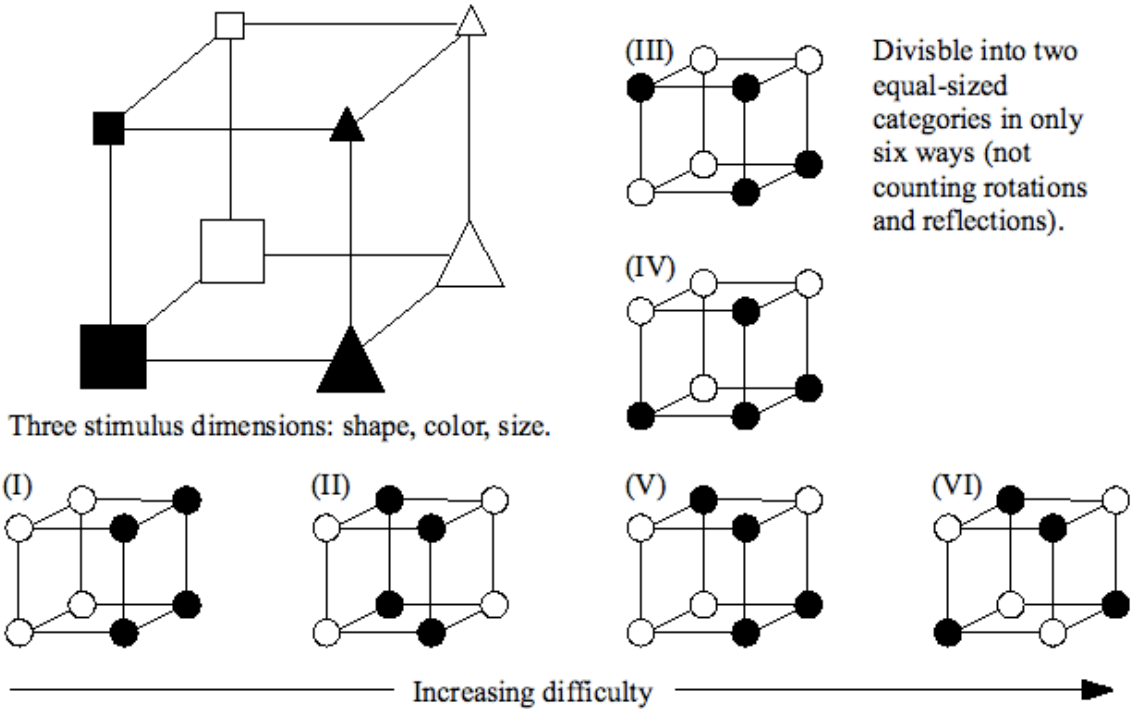
Coefficient	HH	VV	HB	PV
(Intercept)	0.274	0.157	–0.099	–0.211
Studied XY	0.716 *	0.736 *	0.495 .	0.484 .
$V_1 = V_2$ *or* $C_1 = C_2$	–0.259	–0.480 *	—	—
2nd half	–0.278	0.022	0.104	0.169
Studied XY × *2nd half*	–0.059	–0.540	–0.583	–0.428
XY-nonconforming	0.101	0.271 *	–0.115	0.211 .
1st in pair	0.465 **	0.468 **	0.455 **	0.493 **

(8) Summary: Phonological patterns are easier to acquire in the lab when they can be characterized by a single feature, either between or within stimuli.

3 Simplicity bias in non-linguistic category learning

(9) Phonological learning is a kind of category learning (the learner acquires the ability to classify utterances as legal, illegal, or in between). Does category learning in phonology work like category learning in other domains?

(10) Single-feature categories are easier to learn in non-linguistic contexts too (Shepard et al., 1961; Nosofsky et al., 1994; Ashby et al., 1999). There is a hierarchy of difficulty from there on up. Example from Nosofsky et al. (1994) (figures redrawn from there and Griffiths et al. (2008)):



(11) Compare Kuo (2009): Acquisition of new onset phonotactics by Mandarin Chinese speakers. Mandarin allows syllables of the form Consonant-Glide-Vowel-(X); in experiment, glide (j/w) depended on one or more properties of the initial consonant. Results (pooled across test blocks): Proportion correct, significance by t-test vs. chance (=0.50):

Condition	Shepard type	Old		New	
PoA	II	0.68	***	0.62	***
Aspiration	II	0.67	***	0.59	***
Both	VI	0.58	*	0.51	n.s.

(12) Non-linguistic parallels to syntagmatic simplicity bias are much harder to find, primarily because non-linguistic stimuli don't often have multiple instances of a single feature Medin et al. (1982); Billman and Knutson (1996).

4 Simplicity bias and the delta rule

(13) Summary so far: Phonotactic learning is subject to inductive bias favoring paradigmatically- and syntagmatically-simpler patterns. The paradigmatic-simplicity bias, at least, is shared with non-linguistic learning.

(14) There is a long history in phonology of proposed grammatical biases having to do with featural complexity:

a. Preference for rules or constraints involving fewer representational elements (Chomsky and Halle, 1968; Clements, 1995; Sagey, 1990; Clements and Hume, 1995; Gordon, 2004).

b. Preference for within-tier over cross-tier dependencies (Goldsmith, 1976; McCarthy, 1981; Newport and Aslin, 2004).

c. Preference for dependencies between featurally-similar units (Frisch et al., 2004; Rose and Walker, 2004; Onnis et al., 2005).

d. Hard upper limit on representational elements in rule or constraint (Hayes, 1999; Hayes and Steriade, 2004; Hayes and Wilson, 2008)

(15) These biases are normally *imposed* on the learner as heuristics used in choosing between grammars. Is there some way to *derive* them instead? Let's see:

(16) In classical Optimality Theory (Prince and Smolensky, 1993), constraints are ranked. There are alternatives in which constraint dominance is instead expressed by a real number: Harmonic Grammar (Legendre et al., 1990); Stochastic OT (Boersma, 1997); Maximum Entropy grammar (Goldwater and Johnson, 2003).

(17) There are incremental learning algorithms for StOT (Boersma, 1997; Boersma and Hayes, 2001), HG (Pater, 2008; Boersma and Pater, 2008), and ME (Jäger, xxxx) which change the constraint weights only when the learner makes an error, whereupon the "Delta Rule" is applied to derive the new weights $\overrightarrow{r}'$:

$$\overrightarrow{r}'_j = \overrightarrow{r}_j + \eta \cdot (c_j(i, o_{train}) - c_j(i, o))$$

(18) Because $\overrightarrow{r}$ changes only when the learner makes an error, and because the Delta Rule strengthens or weakens constraints in proportion to their contribution to the error, *constraints prosper when they explain data which other constraints don't explain.*

I.e., constraints *compete* to explain the training data, and success is rewarded with influence.

(19) Observation: If constraints are stated in terms of phonetic features, then featurally-simpler patterns are supported by more-general constraints (constraints which apply to more stimuli), whereas featurally complex patterns are supported by parochial constraints (Pater et al., 2008).

Simple: [ptk] vs. [bdg]

	*[+voice]	*b	*d	*g
→p →t →k				
b	–1	–1		
d	–1		–1	
g	–1			–1

Complex: [pdk] vs. [btg]

	*[+voice]	*b	*t	*g
→p →d →k	–1			
b	–1	–1		
t			–1	
g	–1			–1

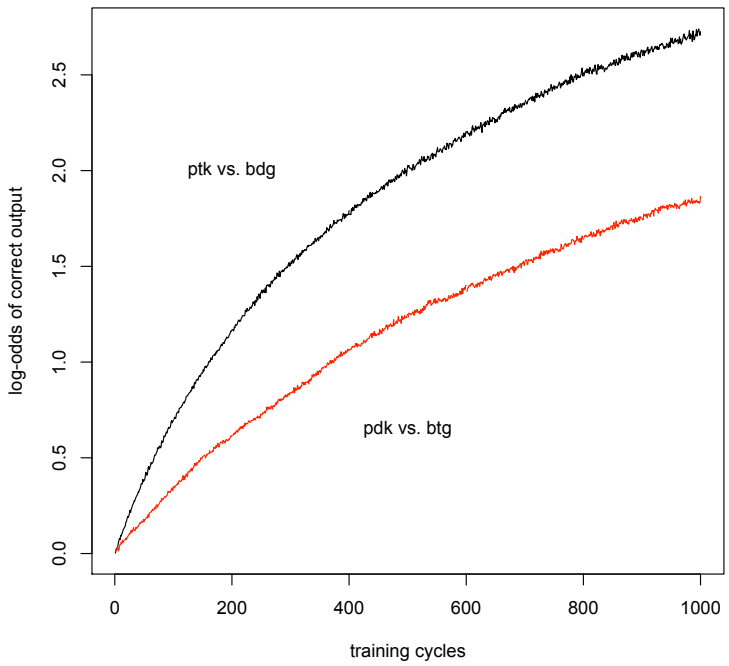
Simple: Voice-voice

	AGREE-[voi]	*[–voi] [+voi]	*[+voi] [–voi]
→pəp →bəb			
pəb	–1	–1	
bəp	–1		–1

Complex: Height-voice

	*[–voi] [+voi]	*[+voi] [–voi]
→æp →ɪb		
æb	–1	
ɪp		–1

(20) Example: [p t k] vs [b d g] is learned faster than [p d k] vs. [b t g] (simulation using ME learner in Praat (Boersma and Weenink, 2010)):



(21) The Delta Rule also plays a major role in models of

a. Human category learning: Category cues (e.g., "has fever") instead of constraint violations (the configural cue model (Gluck and Bower, 1988a,b), ALCOVE (Kruschke, 1992), SUSTAIN (Love et al., 2004)).

b. Classical conditioning in non-humans: Elements of a compound stimulus (e.g., "blue light") instead of constraint violations (Rescorla and Wagner, 1972).

In these models, it does the same thing it does in HG-GLA, viz., strengthen associations between a category and a cue when the cue makes an unexpected correct prediction.

It thus provides a link, and a reason for the commonalities, between phonological learning and other kinds of learning.

---

5 Constraint induction and the Subtree Schema

(22) If constraint weights are learned using the Delta Rule, then the problem of why simpler patterns are easier reduces to that of why the simpler patterns are supported by more-general constraints (= more-valid cues).

E.g., the theory needs to insure that there is a single constraint *[+voice], but not a single constraint *[b ∨ t ∨ g].

(23) We can't assume that the human or model learner comes into the experimental condition with the right constraint set already, since the simplicity advantage does not seem to depend on the specific content of the features.

(24) ⇒ Learner needs to induce constraints from data, but do so in such a way that syntagmatically- and paradigmatically-simpler patterns end up with more-general constraints.

(25) Current constraint-induction models use constraint schemas which are too restrictive for this problem:

a. No syntagmatic feature variables—can't express iff/xor constraints like AGREE or OCP (Gildea and Jurafsky, 1995; Albright and Hayes, 2002; Heinz, 2007; Hayes and Wilson, 2008).

   ⇒ No general constraints for syntagmatically-simple patterns like [α high][α high]. *(But see Sunday's presentation by C. Wilson.)*

b. Practical limits on constraint complexity in order to make exhaustive search feasible, causing difficulties with non-adjacent dependencies (see discussion in Hayes and Wilson (2008, 6.2))

(26) Preview of alternative approach:

a. Constraint schema uses representational (sub-)tree to describe locus of violation/satisfaction. ⇒ Every representation is itself a constraint (Burzio, 1999).

b. Variables for tree nodes allow within-stimulus dependencies, e.g., [α Place]...[α Place].

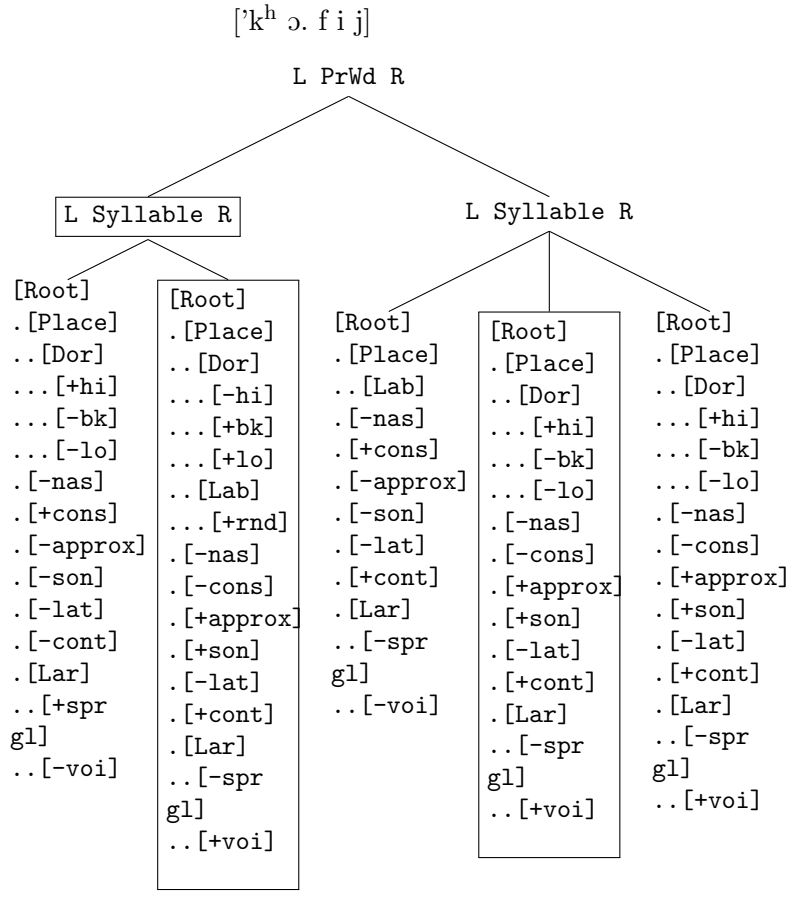
c. Resulting infinite constraint space is searched *non-exhaustively* using evolutionary algorithm (e.g., Eiben & Smith 2003).

d. Fitness determined by Delta Rule; numerosity plays role of weight.

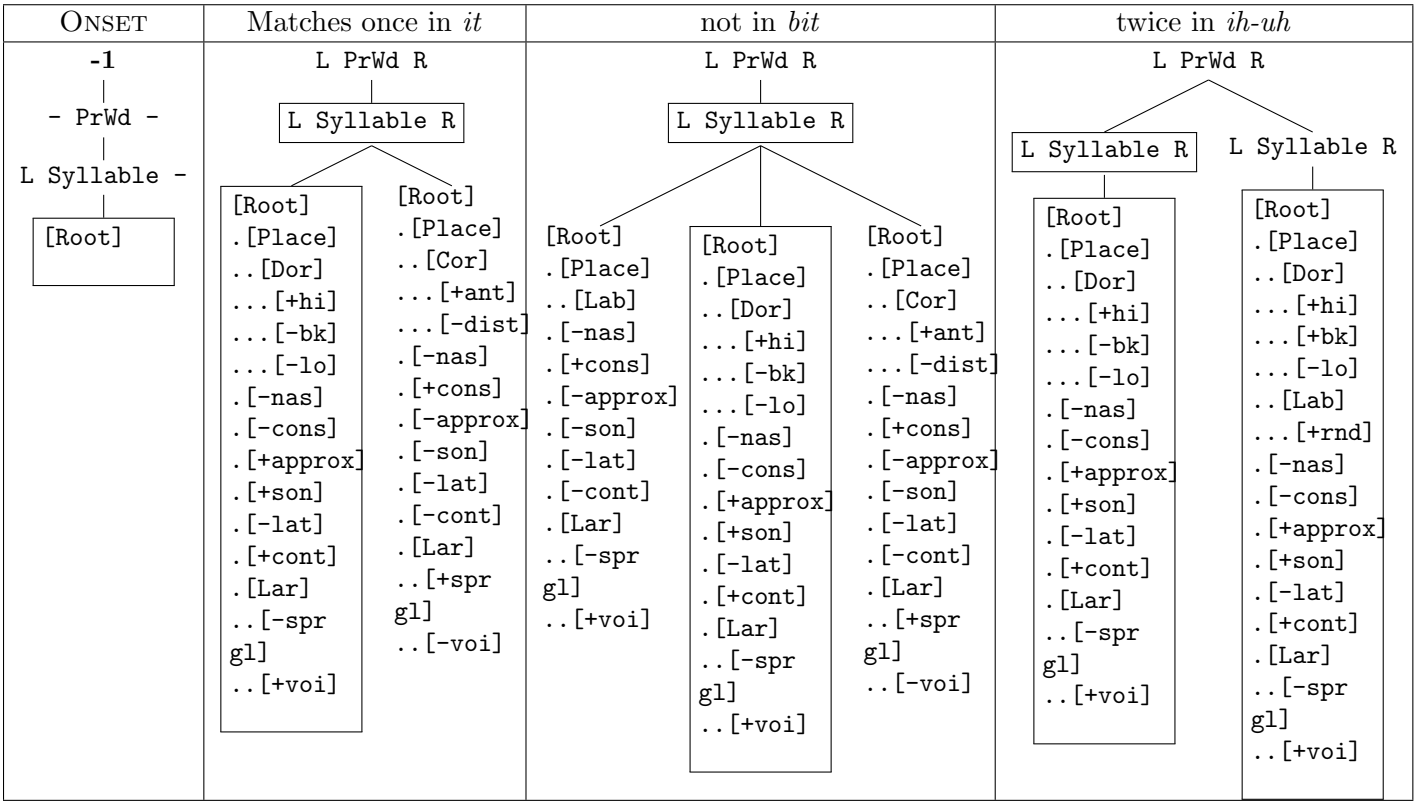
e. Induction and "weighting" of constraints happen simultaneously (Hayes and Wilson, 2008).

5.1 Subtree Schema

(27) Representational system based on Gussenhoven and Jacobs (2005, Ch. 5) (omits onset/nucleus/coda, feet, and moras). Example: *coffee*. Heads are enclosed in a box.

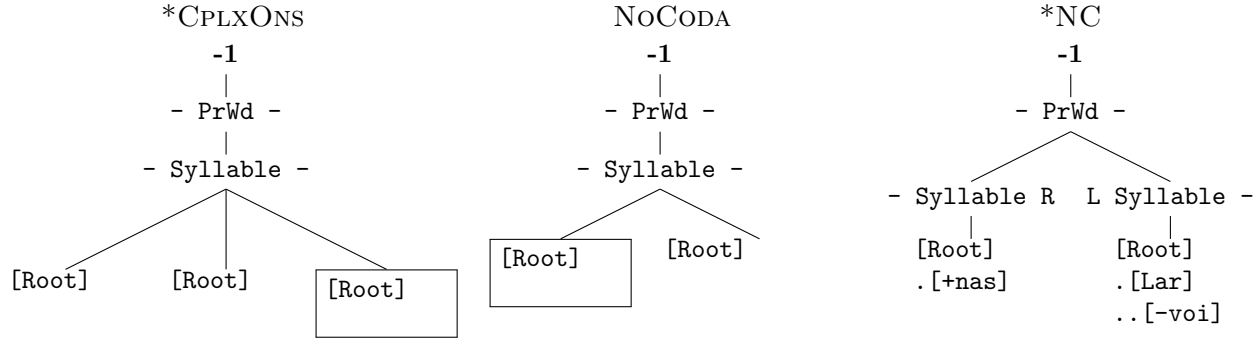


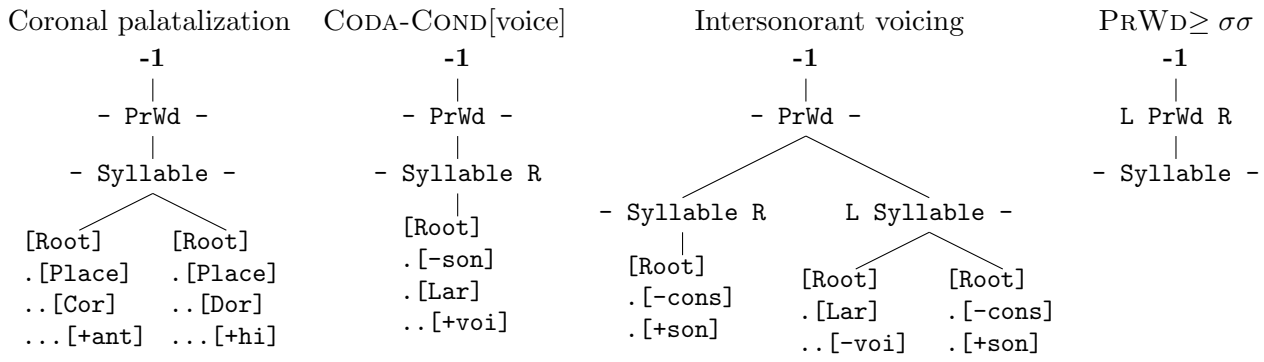
(28) A constraint is just a representation, rooted at a PrWd, with a +1 or –1 to indicate whether it rewards or punishes each match. Here's Onset, à la Smith (2006):



(29) ⇒ Less-stringent constraints match in more-stringent constraints. Fully-specified representations are the most-stringent constraints.

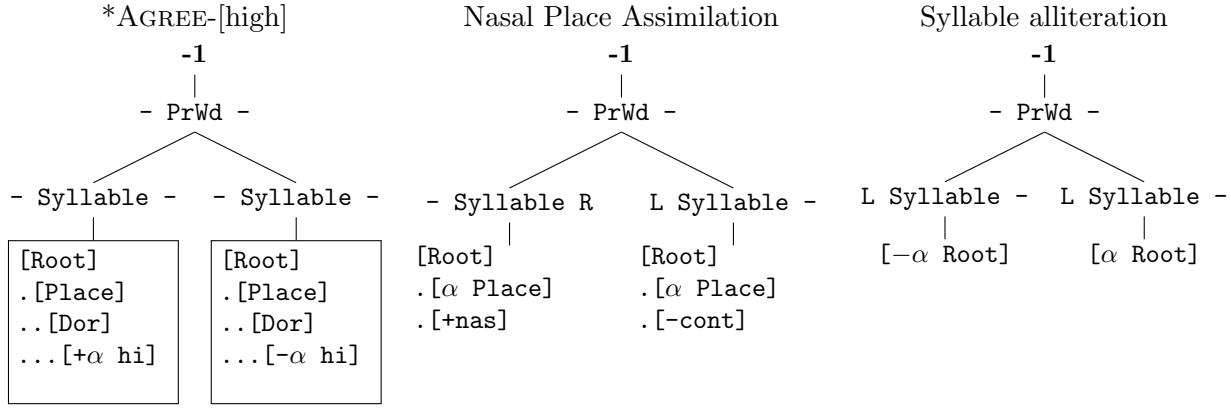
(30) Here are a few familiar constraints.





(31) The value of a Greek-letter variable is an entire node and all of its dependents. E.g., [α Cor]...[α Cor] only matches two Coronal nodes whose dependents, if any, agree in every respect. Consequently, a configuration like [α Cor]...[α Lar] never matches anything.

(32) Some constraints with variables:



5.2 The Delta Rule in constraint-set evolution

(33) Subtree schema permits general constraints only for simple patterns, setting learner up to acquire final ranking faster. But it has to acquire the constraint set first!

(34) A standard solution for searching huge design space is non-exhaustive search by evolutionary algorithm (e.g., Eiben and Smith, 2003).

(35) Harmonic Grammar (Legendre et al., 1990) is "Optimality Theory with weights"—the relative importance of constraints is expressed as weighting, rather than ranking, and the weighted sum of violations determines the winner.

We'll modify this so that all constraints have weight 1, but constraints can be repeated:

ORIGINAL

/atra/	NoCoda 2.0	*CplxOns 1.0	
→a.tra		–1	–1.0
at.ra	–1		–2.0

MODIFIED

/atra/	NoCoda 1.0	NoCoda 1.0	*CplxOns 1.0	
→a.tra			–1	–1.0
at.ra	–1	–1		–2.0

(36) Harmonic Grammar has an on-line learning algorithm, HG-GLA, which adjusts the weights in response to errors (Boersma and Pater, 2008). The algorithm is based on the Delta Rule. If the margin of harmony between outputs o_1 and o_2 should be $\geq t$, but is actually a smaller quantity a, the weights are altered slightly to weaken the o_2-favoring constraints and strengthen the o_1-favoring ones:

$$\Delta w_i = \eta \cdot (C_i(o_1) - C_i(o_2)) \cdot (t - a)$$

where η is adjustable to set the learning rate. (Boersma and Pater implement the margin-of-harmony part in a different way, using "noisy HG".)

(37) Modify the learning algorithm to work with multiplicities rather than weights:

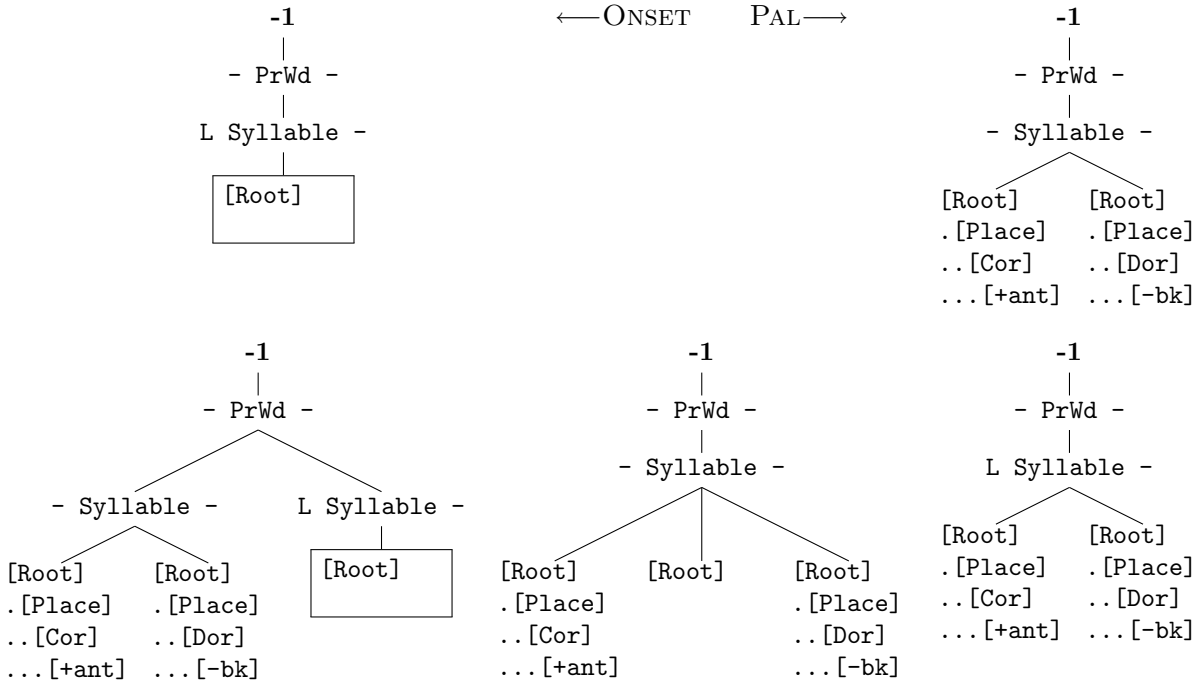
a. Δw_i is interpreted as probability of reproduction (if positive) or deletion (if negative).

b. Number of constraints is a large fixed number; population is controlled by random deletion (similar to "weight decay" in the perceptron; see Hastie et al. 2001, §3.4.3).

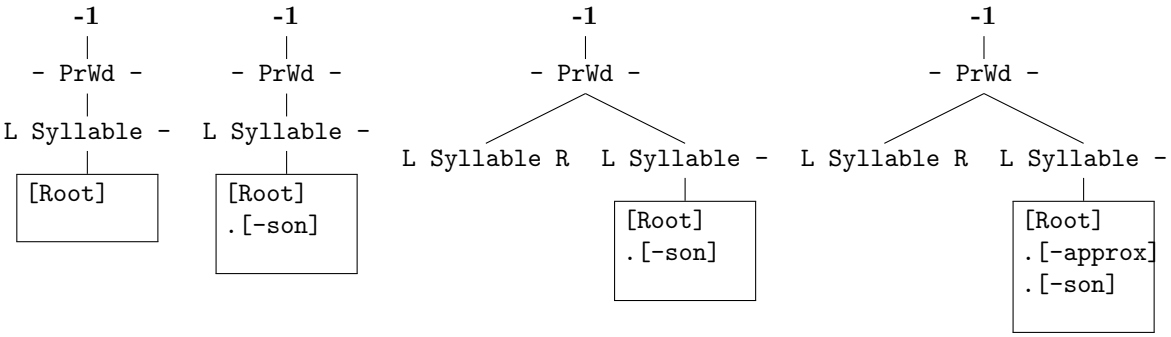
If no novel constraints are introduced during learning, the breeding-and-selection algorithm implements a penalized version of HG-GLA, and learns as if it were an ordinary HG learner.

(38) Now we let the constraints reproduce *with variation*, so that the Delta Rule acts as a selective force in an evolutionary algorithm. Variation comes about in two ways:

a. Breeding with other constraints to yield offspring that randomly combine features of either. The breeding algorithm is recursive; when two nodes are bred, their dependents are randomly aligned and bred too. Example: Offspring of Onset and Pal:



b. The offspring then undergoes random mutation. Mutation is recursive (when a node is exposed to the hazard, so are its dependents). Example: Successive mutations of Onset.



(39) /Input/ on each trial is a random positive stimulus. Candidate [outputs] are that item (the fully-faithful candidate) and one random negative stimulus.

(40) On error, all constraints favoring correct response get chance to breed. More-general constraints are relevant on more trials, and so breed faster when correct.

(41) Prior art: Evolutionary algorithms applied to evolving receptive fields for inputs to single-layer perceptron (Nakano et al., 1995). Applied to evolving tree structures (Cramer, 1985; Koza, 1989).

This model is most similar to the configural-cue model (Gluck and Bower, 1988a, Figure 11), except that it selectively evolves input nodes with successful receptive fields, rather than hard-wiring in all possibilities.

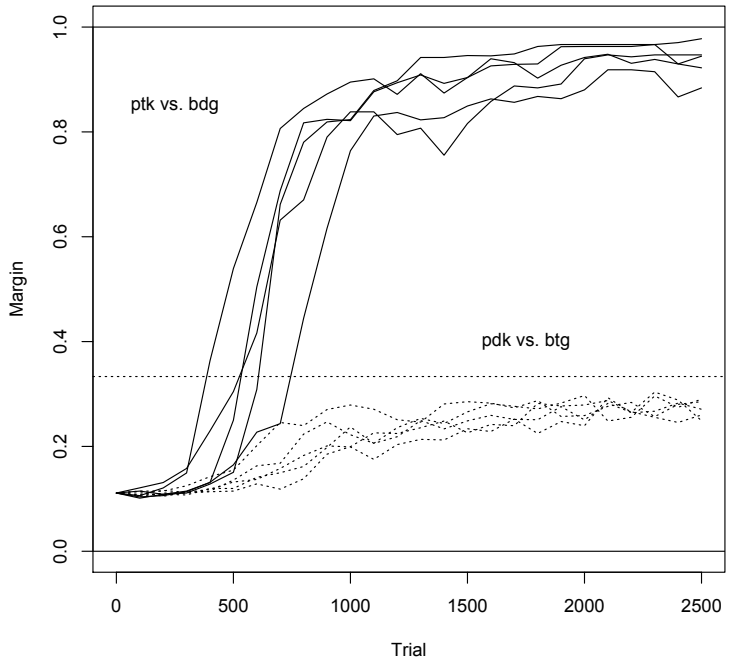
5.3 Simulations

(42) Predictions for supervised discrimination learning:

Simple: ptk vs. bdg	Complex: pdk vs. btg
pɪb tɪd kɪg pæb tæd kæg pʊb tʊd kʊg *vs.* bɪp dɪt gɪk bæp dæt gæk bʊp dʊt gʊk	pɪb dɪt kɪg pæb dæt kæg pʊb dʊt kʊg *vs.* bɪp tɪd gɪk bæp tæd gæk bʊp tʊd gʊk
Prediction: Easier	Prediction: Harder

(43) Initial constraint population: 10 copies of each positive training item as a positive constraint (rewards matches), and 10 of each negative training item as a negative constraint (punishes matches). Similar to Albright and Hayes (2002) Minimal Generalization Learner—start with the data and simplify it.

(44) Simulation results, 5 runs in each condition. Vertical axis is average harmony difference between positive and negative stimuli, as a proportion of the target margin of separation. (This would go through a squashing function to model response probabilities.)



The final constraint population in Condition 1 is dominated by a single constraint type that prefers voiceless onsets, voiced codas, or both; that in Condition 2 is a roughly equal mix of constraints preferring specific segments.

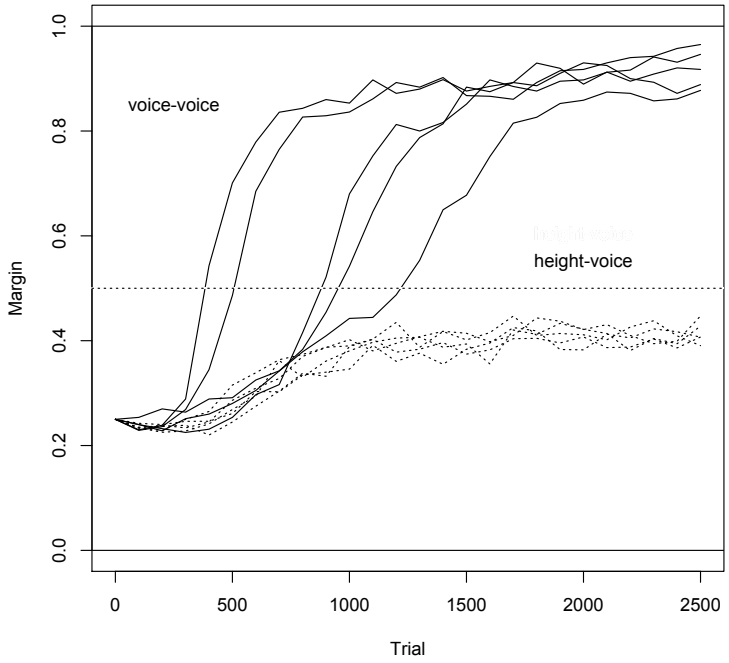
(45) The theoretical maximum performance is never reached in either condition, because a certain proportion of the population always consists of lower-quality mutants.

(46) What about syntagmatic simplicity bias?

Simple: [α voice]... [α voice]	Complex: [α high]... [α voice]
pæp pɪp bæb bɪb *vs.* bæp bɪp pæb pɪb	pæp bæp pɪb bɪb *vs.* pɪp bɪp pæb bæb
Prediction: Easier	Prediction: Harder

(47) Initial constraint population: 25 copies of each positive and each negative training item as positive and negative constraints.

(48) Simulation results:



The final constraint population in Condition 1 is dominated by a single constraint type that prefers voicing (or spread-glottis) agreement between the initial and final consonant. In Condition 2, the lion's share is divided evenly between two constraint types, one favoring [+high][+voiced] and the other favoring [–high][–voiced].

(49) ⇒ Model does in fact show both paradigmatic and syntagmatic simplicity bias, in that learning of the simpler patterns is faster and reaches a higher final level of performance. It does this because

a. It evolves its constraint set according to the Delta Rule, favoring more-general solutions when they are available.

b. The constraint schema determines how general a solution is possible for any given pattern.

(i) Use of representational subtrees as constraints is root cause of paradigmatic simplicity bias.

(ii) Incompatibility of Greek-letter variables in different feature contexts (e.g., [α Place][α Laryngeal]) is root cause of syntagmatic simplicity bias.

c. The evolutionary algorithm allows the (infinite!) space of possible constraints to be searched relatively efficiently.

d. The pdk/btg and HV simulations show the development of a "constraint ecology" in which different constraint types coexist stably, each one making just enough mistakes to feed the others without starving itself.

---

6 Discussion

(50) Empirical problem: Paradigmatic and syntagmatic simplicity bias, in phonology and elsewhere.

(51) Core of solution: The Delta Rule facilitates acquisition of patterns which are supported by more-general constraints (Pater et al., 2008). Link between phonological and non-phonological learning; applies across wide range of phonological and non-phonological learning models.

(52) Main beneficial features of the subtree/delta-rule model

a. Constraints-as-subtrees

(i) No arbitrary limit on constraint size; (some) long-distance dependencies handled straightforwardly.

(ii) Integration of prosodic and Feature-Geometric structure (in both constraints and representations)

i. Restricts constraint space

ii. Induces mutation and breeding algorithms

b. Non-exhaustive search using evolutionary algorithm; fitness determined by Delta Rule.

(i) Connection to non-linguistic category learning.

(ii) Simplicity/generality preference *emerges* from the way the acquisition mechanisms work, rather than being imposed from outside as a grammar-selection criterion (Hale and Reiss, 2000, fn. 8, p. 164).

(53) No other model at present explains syntagmatic simplicity bias, because none of them represent the predicate "two instances of the same feature". This is true for both phonological (Gildea and Jurafsky, 1995; Albright and Hayes, 2002; Heinz, 2007; Hayes and Wilson, 2008) and psychological (Rescorla and Wagner, 1972; Gluck and Bower, 1988a,b; Kruschke, 1992; Love et al., 2004) models.

How hard would it be to modify them to do so?

(54) ⇒ Explanatory focus shifts to

a. Constraint schemas, generation, and testing (Boersma, 1998; Hayes, 1999; Smith, 2002, 2004; Boersma and Pater, 2007). E.g.,

(i) Can Constraint X be expressed in a single constraint, or do we need more than one? (As in the syntagmatic and paradigmatic exx. above.)

(ii) How many extensionally-equivalent ways are there to express Constraint X? The more there are, the sooner one will be found!

b. If constraints are representations, convergence of constraint schemas with representational schemas like Feature Geometry.

c. If space is searched by mutation and selection, considerations of mutation algorithm. Mutation distance between constraints also depends on choice of schema.

(55) Most serious problems with model:

a. Requires explicit negative data (losing candidates) during training.

(i) Possible solution for lab situation: Use L_1 lexicon as negative data.

(ii) However, this doesn't work for the HH/HV experiment: The positive training items differ from the run of the English lexicon in too many ways. The learner finds a near-perfect solution that is almost never the "right" one.

(iii) Also, that approach can't be used when what is being learned is the L_1 itself.

b. Delta Rule leads to low long-term constraint diversity, since successful general constraints block acquisition of more-specific constraints. ⇒ Won't find niche generalizations like the more-thorough searches of Albright and Hayes (2006); Hayes and Wilson (2008). Do we need exhaustive search after all?

c. Fixed population size imposes artificial upper limit on performance (no overlearning).

d. No faithfulness constraints.

e. Can't ignore prosodic structure (see examples in (30) above).

(56) Empirical questions:

a. Do the Shepard et al. (1961) results on paradigmatic simplicity bias hold for phonological stimuli? If not, when do they break down?

b. Are syntagmatic-simplicity-bias effects unique to phonology?

c. Do the standard Rescorla-Wagner-type effects turn up in phonology? E.g., does learning to solve a problem using one constraint block subsequent discovery of another?

References

Albright, A. and B. Hayes (2002). Modelling English past tense intuitions with minimal generalization. In M. Maxwell (Ed.), *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology, Philadelphia, July 2002*, pp. ??–?? Association for Computational Linguistics.

Albright, A. and B. Hayes (2006). Modeling productivity with the Gradual Learning Algorithm: the problem of accidentally exceptionless generalizations. MS, Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Ashby, F. G., S. Queller, and P. Berretty (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics 61*(6), 1178–1199.

Becker, M., N. Ketrez, and A. Nevins (2007). Where and why to ignore lexical patterns in Turkish obstruent alternations. Handout, 81st annual meeting of the Linguistic Society of America, Anaheim, California.

Bell, A. (1971). Some patterns of the occurrence and formation of syllabic structure. *Working Papers on Language Universals 6*, 23–138.

Billman, D. and J. Knutson (1996). Unsupervised concept learning and value systematicity: a complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*(2), 458–475.

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

Boersma, P. (1998). *Functional Phonology: formalizing the interactions between articulatory and perceptual drives*. Ph. D. thesis, University of Amsterdam.

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

Boersma, P. and J. Pater (2007, October). Constructing constraints from language data: the case of Canadian English diphthongs. Handout, NELS 38, University of Ottawa.

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

Boersma, P. and D. Weenink (2010). PRAAT Version 5.1.31. Software, www.praat.org.

Burzio, L. (1999). Surface-to-surface morphology: when your representations turn into constraints. MS, Department of Cognitive Science, Johns Hopkins University. ROA-341.

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

Clements, G. N. (1995). The geometry of phonological features. In J. A. Goldsmith (Ed.), *Phonological theory: the essential readings*, pp. 201–223. Malden: Blackwell.

Clements, G. N. and E. V. Hume (1995). The internal organization of speech sounds. In J. A. Goldsmith (Ed.), *The handbook of phonological theory*, Chapter 7, pp. 245–306. Boston: Blackwell.

Cramer, N. L. (1985). A representation for the adaptive generation of simple sequential programs. In J. Grefenstette (Ed.), *Proceedings of the First International Conference on Genetic Algorithms*, pp. 183–187.

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

Eiben, A. E. and J. E. Smith (2003). *Introduction to evolutionary computing*. Berlin: Springer.

Frisch, S., J. B. Pierrehumbert, and M. B. Broe (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory 22*(1), 179–228.

Gildea, D. and D. Jurafsky (1995). Automatic induction of finite-state transducers for simple phonological rules. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics (ACL-95), Cambridge, Massachusetts*, pp. 9–15. Association for Computational Linguistics.

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.

Goldsmith, J. A. (1976). *Autosegmental phonology*. Ph. D. thesis, Massachusetts Institute of Technology.

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.

Gordon, M. (2004). Syllable weight. In B. Hayes, R. Kirchner, and D. Steriade (Eds.), *Phonetically-based phonology*, pp. 277–312. Cambridge, England: Cambridge University Press.

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., 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.

Gussenhoven, C. and H. Jacobs (2005). *Understanding phonology* (2nd ed.). Understanding Language Series. London: Hodder Arnold.

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning: data mining, inference, and prediction*. Berlin: Springer Verlag.

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 D. Steriade (2004). The phonetic bases of phonological Markedness. In B. Hayes, R. Kirchner, and D. Steriade (Eds.), *Phonetically-based phonology*, Chapter 1, pp. 1–33. Cambridge, England: Cambridge University Press.

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

Healy, A. F. and A. G. Levitt (1980). Accessibility of the voicing distinction for learning phonological rules. *Memory and Cognition 8*(2), 107–114.

Heinz, J. (2007). Learning phonotactic grammars from surface forms. In D. Baumer, D. Montero, and M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, Somerville, pp. 186–194. Cascadilla.

Jäger, G. (xxxx). 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.

Koza, J. R. (1989). Hierarchical genetic algorithms operating on populations of computer programs. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Volume 1, San Mateo, California, pp. 768–774. Morgan Kaufmann.

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.

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.

Legendre, G., Y. Miyata, and P. Smolensky (1990). Can connectionism contribute to syntax? Harmonic Grammar, with an application. In M. Ziolkowski, M. Noske, and K. Deaton (Eds.), *Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 237–252. Chicago Linguistic Society.

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

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.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry 12*, 373–418.

Medin, D. L., M. W. Altom, S. M. Edelson, and D. Freko (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Human Learning and Memory 7*, 355–368.

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

Moreton, E. (2008b). Modelling modularity bias in phonological pattern learning. In N. Abner, J. Bishop, and K. Ryan (Eds.), *Proceedings of the 27th meeting of the West Coast Conference on Formal Linguistics*

(WCCFL), pp. 1–16.
Myers, S. (2002). Gaps in factorial typology: The case of voicing in consonant clusters. MS, University of Texas, Austin.
Nakano, K., H. Hiraki, and S. Ikeda (1995). A learning machine that evolves. In *Proceedings of ICEC–95*, pp. 808–813.
Newport, E. and R. N. Aslin (2004). Learning at a distance i: statistical learning of non-adjacent dependencies. *Cognitive Psychology 48*, 127–162.
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.
Onnis, L., K. Richmond, and N. Chater (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language 53*, 225–237.
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.
Peperkamp, S., K. Skoruppa, and E. Dupoux (2006). The role of phonetic naturalness in phonological rule acquisition. In D. Bamman, T. Magnitskaia, and C. Zoller (Eds.), *Papers from the 30th Boston University Conference on Language Development (BUCLD 30)*, pp. 464–475.
Prince, A. and P. Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Department of Linguistics, Rutgers University.
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.
Pycha, A., E. Shin, and R. Shosted (2007). Directionality of assimilation in consonant clusters: an experimental approach. MS, Department of Linguistics, University of California, Berkeley.
Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black and W. F. Prokasy (Eds.), *Classical conditioning*, Volume II: Current research and theory. New York: Appleton–Century–Crofts.
Rose, S. and R. Walker (2004). A typology of consonant agreement as correspondence. *Language 80*(3), 475–531.
Saffran, J. R. and E. D. Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology 39*(3), 484–494.
Sagey, E. (1990). *The representation of features in non-linear phonology: the Articulator Node Hierarchy*. New York: Garland.
Seidl, A. and E. Buckley (2005). On the learning of arbitrary phonological rules. *Language Learning and Development 1*(3 & 4), 289–316.
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. L. (2002). *Phonological augmentation in prominent positions*. Ph. D. thesis, University of Massachusetts, Amherst.
Smith, J. L. (2004). Making constraints positional: towards a compositional model of con. *Lingua 114*(2), 1433–1464.
Smith, J. L. (2006). Representational complexity in syllable structure and its consequences for Gen and Con. MS, Department of Linguistics, University of North Carolina, Chapel Hill. ROA-800.
Wilson, C. (2003, January). Analytic bias in artificial phonology learning: consonant harmony vs. random alternation. Handout from presentation at the Workshop on Markedness and the Lexicon, Massachusetts Institute of Technology.