Constraint breeding and incremental learning

E. Moreton
UNC-Chapel Hill

(1) At least *some* constraints don't come for free in the initial state; the learner has to induce them from phonological data. Cogent examples: constraints which...

a. refer to specific lexemes (e.g., McCarthy and Prince 1993; Fukazawa 1999; Pater 2000; Ota 2004; Pater 2007; Coetzee and Pater 2008; Pater 2009; Becker 2009)
b. refer to specific lexical strata, inflectional paradigms, or other language-particular classes (e.g., Benua 1997; Alderete 1999; Ito and Mester 2001; Flack 2007; Inkelas 2008)
c. refer to phonetically arbitrary segment classes that do not recur across languages (e.g., Bach and Harms 1972; Anderson 1981; Buckley 2000) or enforce idiosyncratic requirements (e.g., Prince and Smolensky 1993, 101).

(Separate issue, not addressed here: constraint induction from phonetics ("inductive grounding", Hayes 1999).)

(2) How and when are phonological markedness constraints induced? Proposals in the OT/HG literature fall into two main categories:

a. *Exhaustive search*: Learner checks all of a set of possible constraints, keeping those that best satisfy criteria (Hayes and Wilson, 2008; Wilson and Gallagher, 2018).
b. *Error-patching*: Learner identifies a particular error (or class of errors) and makes a constraint against it (Adriaans and Kager, 2010; Pizzo, 2013; Pater, 2014). With positive constraints the learner can identify correct forms and make constraints that reward them (Boersma and Pater, 2007).

(3) Alternative: *Evolution*. Constraints breed with variation and selection. *Evolutionary Winnow-MaxEnt* architecture: Constraints interact via Max Ent HG, but weights are population sizes, weight update is population growth or shrinkage in response to fitness-based selection, and constraint generation happens via mutation and recombination (Moreton 2010b,a,c).

(4) Contents of this talk:

§1 Max Ent/HG weights as population sizes: micro-constraints in a macro-constraint
§2 Macro-constraint weight update as micro-constraint reproduction. Implements (a variant of) Winnow-2 (Littlestone, 1988), shown here to converge.
§3 Markedness constraints as subtrees of candidates/stimuli
§4 Mutation and recombination of subtree constraints
§5 Model property: Abrupt learning
§6 Model property: Constraints prime related constraints
§7 Discussion: Inductive biases. Limiting behavior approximating other models.

[0]The author is indebted to Brian Hsu, Katya Pertsova, and Jen Smith, Chris Wiesen of the Odum Institute at UNC-CH, and three anonymous SCiL reviewers for comments on previous drafts. The paper that goes with this talk (Moreton, 2019) benefited from audience comments on portions of Sections 3 and 4 at the Workshop on Computational Modelling of Sound Pattern Acquisition (University of Alberta, February 14, 2010), at an MIT departmental colloquium (April 9, 2010), and at the Workshop on Grammar Induction (Cornell University, May 14, 2010). The research was supported in part by NSF BCS 1651105, "Inside phonological learning", to E. Moreton and K. Pertsova. Address for correspondence about this talk: `moreton@unc.edu`.

1 Weights as population sizes in a Max Ent/HG learner

(5) In a Harmonic Grammar framework (Legendre et al., 1990), without changing the harmony of any candidate, we can replace any constraint of weight w with k "micro-constraints", i.e., clones of that constraint, each with weight w/k:

Macro-constraints:	*CpOns				Max			
Weights:	4				3			
Micro-constraints:	*CpOns	*CpOns	*CpOns	*CpOns	Max	Max	Max	
Weights:	1	1	1	1	1	1	1	
/bʃib-dʒu/								
[bʃib.dʒu]	*	*	*	*				$H = -4$
→[ʃib.dʒu]					*	*	*	$H = -3$

For higher resolution, we can replace a constraint of weight w with w/ζ constraints that each give ζ marks (e.g., if we want two decimal places of precision, let $\zeta = 0.01$, and then replace Max of weight 3 with 300 micro-Max's of weight 0.01).

(6) *Luce/MaxEnt choice rule*: Given the experimenter's intended winner x^+ and intended loser x^-, the learner chooses x^+ with a probability that depends on the harmonies of the candidates.

$$\Pr(x^+ \mid x^+, x^-) = \frac{\exp(\sum_{i=1}^{n} x_i^+ w_i)}{\exp(\sum_{i=1}^{n} x_i^+ w_i) + \exp(\sum_{i=1}^{n} x_i^- w_i)}$$

This is the Luce choice rule (Luce, 1959, 23) applied to the exponentiated harmonies, i.e., a conditional Maximum Entropy model (Goldwater and Johnson, 2003; Jäger, 2007; Hayes and Wilson, 2008). (Only two-candidate version discussed here, but generalization to k alternatives is straightforward.)

2 Weight update as reproduction

(7) When an error occurs, each micro-constraint produces an offspring with probability $(1 + \epsilon)^d$, where ϵ is a learning-rate parameter and d is the difference between the winner's and loser's score on that micro-constraint. *Example*: Suppose C is binary (gives either 0 or 1 mark). Then

d	Favors	Expected offspring	Population
-1	loser	$1/(1+\epsilon) < 1$	shrinks
0	neither	1	stays same
$+1$	winner	$1 + \epsilon > 1$	grows

(8) *What algorithm does that implement* in terms of the macro-constraint weights?

Not the Perceptron or GLA or gradient ascent or other algorithms where the update rule is $\Delta w_i = \epsilon(t_i - o_i)$ (Jäger, 2007); those increment or decrement each (macro-)weight by a fixed amount that does not depend on how big the weight already is.

Instead, it implements a variant of the Winnow-2 algorithm of Littlestone (1988). This algorithm was mentioned as a possible HG learning algorithm by Magri (2013). The proposal here differs from Winnow-2 in the following ways:

	Winnow-2	Winnow MaxEnt
Task	yes-no classification	k-alternative forced choice
Response	deterministic	probabilistic
Constraints	binary and positive (0/1)	can be non-binary or negative

(9) *Convergence of Winnow-MaxEnt*: The probabilistic response rule makes Littlestone (1988)'s convergence proof for Winnow-2 inapplicable. But we can show that if a concept is linearly separable in terms of the constraint scores, then Winnow-MaxEnt will learn it to any degree of cumulative accuracy (details in Moreton 2019).

a. *Cumulative average of winner-loser harmony gap (relative to total weight) approaches theoretical maximum.* Suppose that there exist nonnegative weights $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and a $\delta_{\boldsymbol{\mu}}$ such that for every candidate pair (x^+, x^-) in the training/testing set,

$$\sum_{i=1}^{n} \mu_i (x_i^+ - x_i^-) > \delta_{\boldsymbol{\mu}} > 0 \tag{1}$$

Let $A_{\max}$ be the least upper bound on $A_{\boldsymbol{\mu}}$ over all $\boldsymbol{\mu}$. For a given winner-loser pair (x^+, x^-), let $\Sigma^+ = \sum_{i=1}^{n} x_i^+ w_i$ and $\Sigma^- = \sum_{i=1}^{n} x_i^- w_i$, and let $a = (\Sigma^+ - \Sigma^-)/W$.
Then for any $\theta > 0$, there exist ϵ_θ and t_θ such that when Winnow-MaxEnt is run with $\epsilon = \epsilon_\theta$,

$$\frac{1}{t} \cdot \sum_{\tau=0}^{t-1} a(\tau) \geq A_{\max} - \theta \tag{2}$$

for all $t \geq t_\theta$ (where t indexes trials where an update was made, i.e., errors).

b. *Cumulative average log-odds correct grows without bound.* Let $L(t)$ be the cumulative average log-odds of a correct response on Error Trials 1 to $t-1$ (i.e., the log odds of making an error, on each trial where an error was actually made). Then

$$L(t) \geq \frac{\mu_0 d_{\max}}{A_{\max} \epsilon_\theta} \frac{1}{t} \left(\exp\left(\frac{A_{\max} \epsilon_\theta (A_{\max} - \theta)}{d_{\max}} t \right) - 1 \right) \tag{3}$$

where $\mu_0 = \min_i(w_i(0))$, i.e., the smallest initial weight, and $t \geq t_\theta$.

(10) For a fixed θ, the time required to satisfy Inequality 2 is $O(\log n)$, i.e., not very sensitive to the number of (macro-)constraints, which is good news.

These bounds were verified with 10,000 learning simulations using random concepts (see Moreton 2019 for details). They are worst-case bounds, and the simulations (a) came close to the bounds without crossing them, and (b) showed that the median case (for the given sampling method) is a lot better than the bound, meaning that the bounds don't necessarily tell us much about how long things will take in practice.

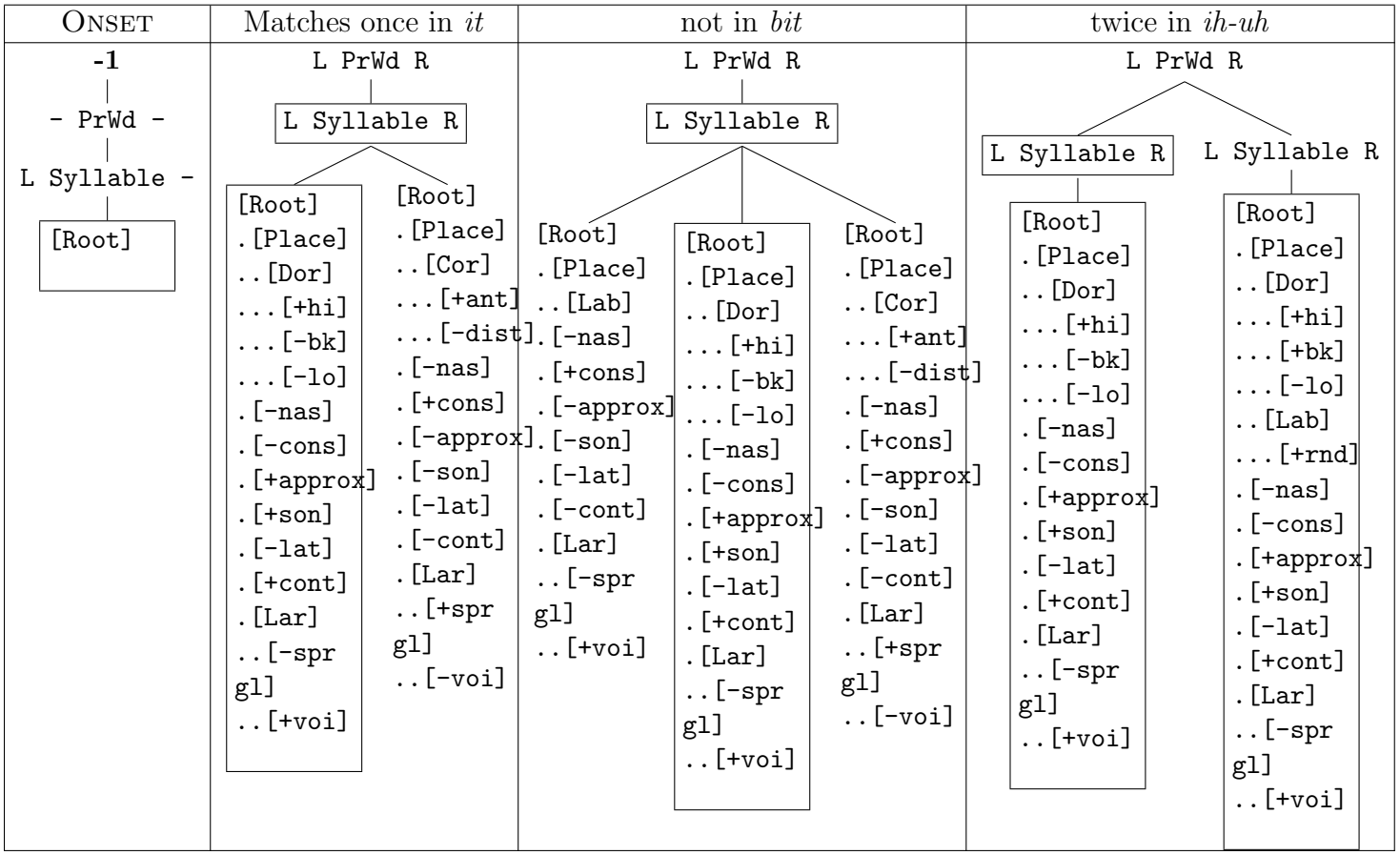
(11) ⇒ Replacing the familiar Perceptron-like algorithms by Winnow MaxEnt won't wreck the learner's ability to find weightings that work.

(12) Weights can grow or shrink quickly, so population explosion and total extinction are both possible. Can be dealt with by randomly cloning or deleting micro-constraints to maintain a fixed population size.

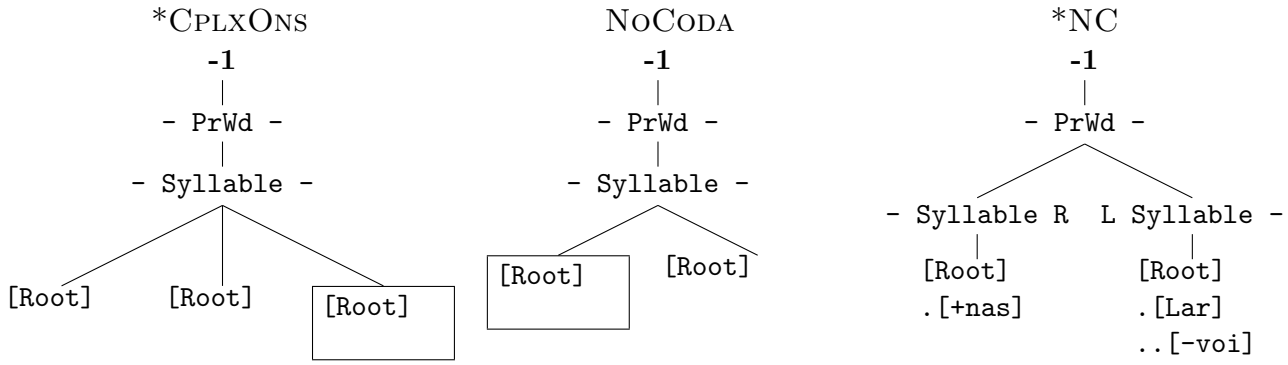
3 Constraints as subtrees of representations

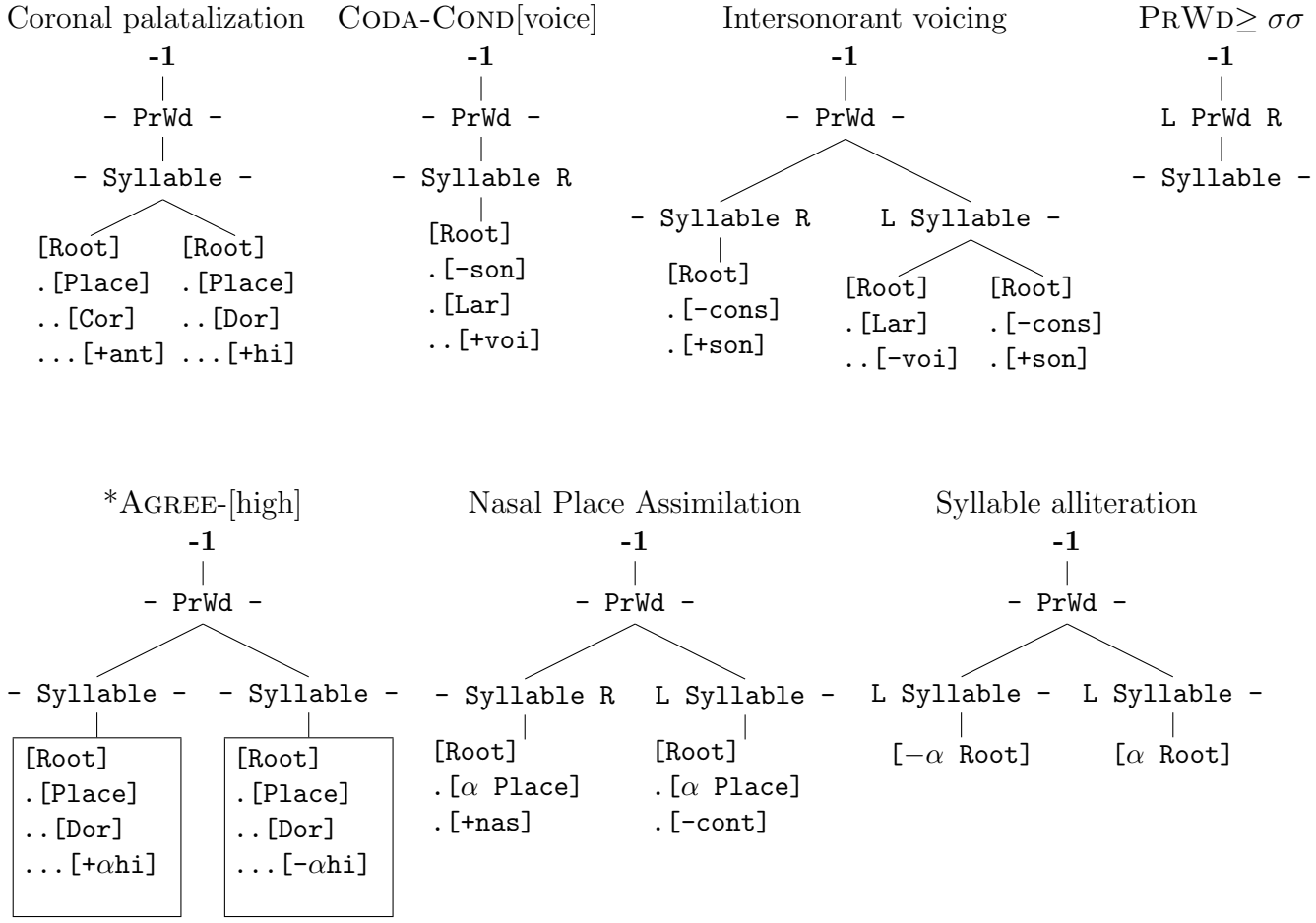
(13) For phonological representations, we can use prosodic and Feature-Geometric trees familiar from existing phonological theory (Goldsmith, 1976; McCarthy, 1981; Sagey, 1990; Clements and Hume, 1995). This example uses a slightly simplified version of the one in Gussenhoven and Jacobs (2005, Ch. 5). The box marks the [head].

(14) A constraint is a representation, rooted at a PrWd, which describes a locus of violation or of satisfaction. Here is ONSET, à la Smith (2006):



(15) Further illustrations of the Subtree Schema:





(16) Properties of the Subtree Schema:

a. Imposes no extra restrictions on markedness constraints beyond those inherited from the Autosegmental/Feature-Geometric representational system.

b. Supports both adjacent and non-adjacent dependencies (e.g., Nasal Place Assimilation and Agree-[high] in 15)

c. Supports lexical exceptions natively. (Continuity between representations and constraints means continuity between grammar and lexicon.)

d. Supports Greek-letter variables for Agree, OCP, reduplication (not discussed here; see Moreton 2010c)

e. Lends itself to recursive recombination and mutation (see next section)

4 Constraint generation as mutation and recombination

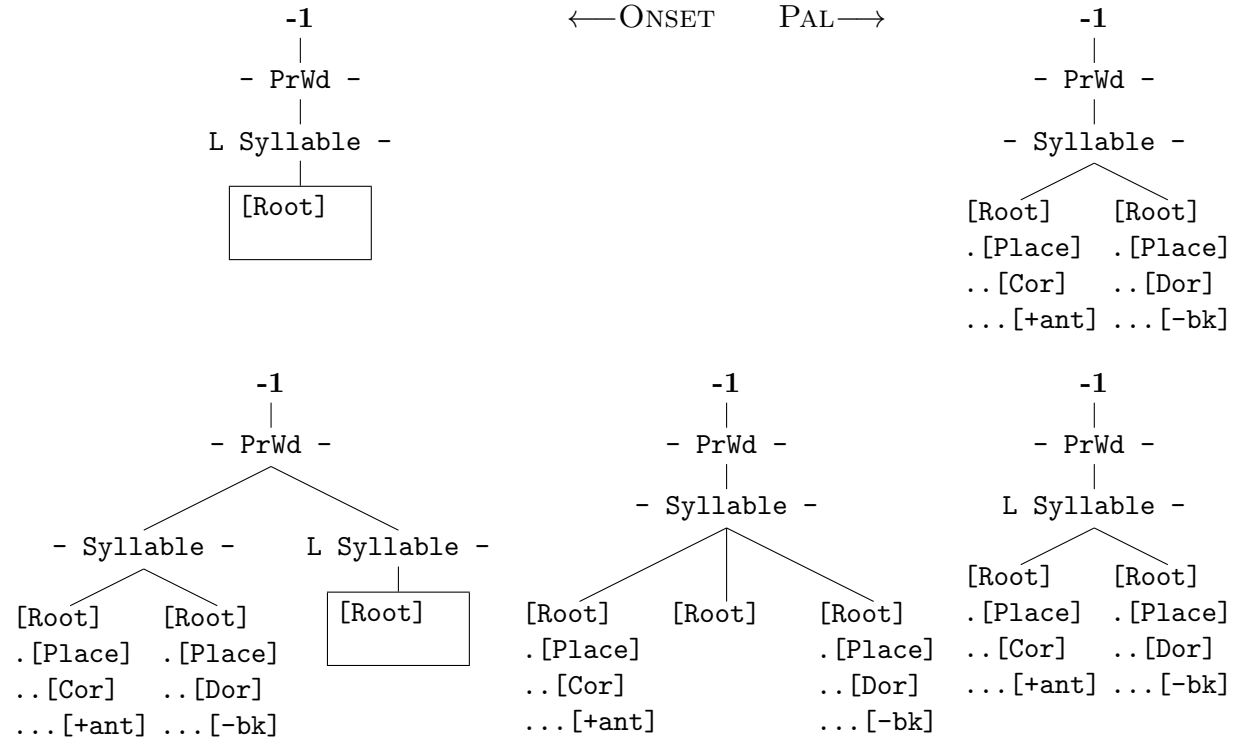
(17) Now we let the micro-constraints reproduce *with variation*, so that the update rule acts as a selective force in an evolutionary algorithm.

a. Evolutionary algorithms have long been applied to problems closely related to the ones addressed here, including evolving receptive fields for inputs to the single-layer perceptron (Nakano et al., 1995) and evolving tree structures (Cramer, 1985; Koza, 1989).

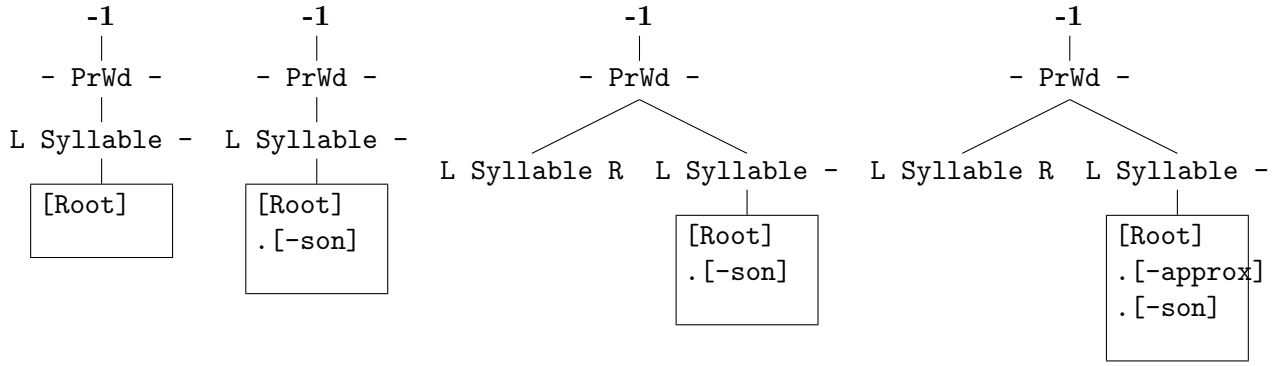
b. Replication of mental representations with variation, recombination, and selection is a leading theory of human creativity in other domains (Simonton, 1999, 2004; Dietrich and Haider, 2015).

(18) Variation comes about in two ways:

a. Recombination 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.



(19) Mutation is insensitive to the data or the state of the model, unlike the error-patching learners cited in (2) above. Human creativity in other domains may or may not work the same way (Campbell, 1960; Simonton, 1999; Dietrich and Haider, 2015).

(20) The combination of Winnow MaxEnt, the Subtree Schema, and evolution allows the model to learn symbolic constraints via a connectionist learning rule.

a. Symbolic constraints avoid two shortcomings that have been pointed out in connectionist intermediate nodes, bounded retinal width (Minsky and Papert, 1969) and inability to generalize to new features (Marcus et al., 1999; Berent et al., 2012).

b. The constraint set does not need to be prespecified, avoiding the combinatorial explosion found in GMECCS (Pater and Moreton, 2012; Moreton et al., 2017).

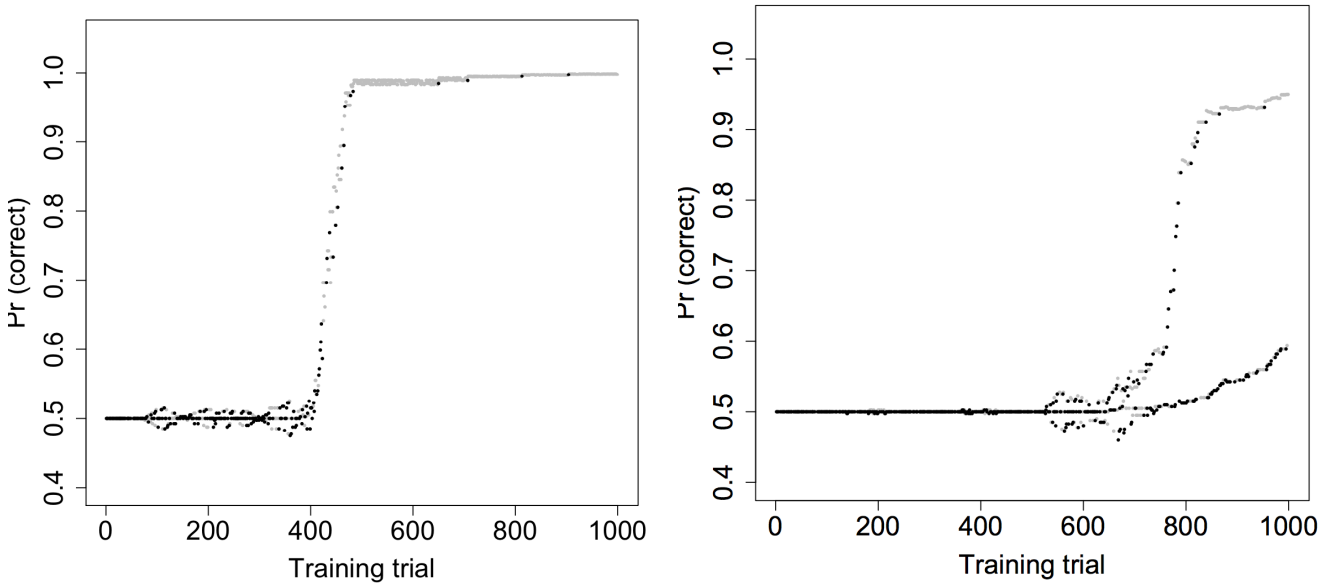
5 Model property: Abruptness

(21) Human phonological learning (in nature and the lab) can be abrupt, in that a long period of apparent stagnation is followed by significant improvement (see refs in Moreton 2018).

(22) Sigmoidal abruptness in an observed learning curve has often been taken as distinguishing "rule-based" learning by serial hypothesis testing from "cue-based" associative learning by gradual weight changes (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al., 2012; Kurtz et al., 2013). The theory is that while the curve is flat, the learner is serially testing and discarding incorrect rule hypotheses, and the jump occurs when the correct rule is found.

(23) Evolutionary Winnow-MaxEnt can show similar behavior while searching for a constraint. Example (from Moreton 2019): Candidates were the initial syllables from the stimuli of Saffran and Thiessen (2003, 494). The positive stimuli (winners) had the form $\{p,t,k\}V\{b,d,g\}$ and the negative stimuli (losers) $\{b,d,g\}V\{p,t,k\}$. $N = 1000$ constraints initialized to `*(L PrWd R)`.

Left panel, fell-swoop constraint is discovered and abruptly solves problem. Right side, parochial constraint (applies only to long-vowel syllables) is discovered first, and delays general solution. (Gray = correct response, black = error.)



(24) Prediction: More abrupt learning for constraints that have to be induced vs. constraints that are given by UG or that were already acquired in L1 (Becker and Tessier, 2011).

6 Model property: Priming and attention

(25) Because new macro-constraints are founded by mutation out of old ones, the Evolutionary Winnow-MaxEnt model predicts that *existing macro-constraints prime discovery of new ones that are similar to them.*

Because high-weighted macro-constraints initiate more mutations, the model predicts that *new macro-constraints tend to be mutants of high-weighted old ones.*

(Let $\pi_{i,j}$ be the probability that a random micro-constraint in C_i will breed a micro-constraint in C_j on the next error trial. Then the probability of discovering C_j from C_i is approximately $w_i\pi_{i,j}$,

if $\pi_{i,j}$ is small. The bigger w_i is, the higher that probability.)

(26) Simple, concrete illustration: Stimulus space is the set of all $(C)V(C) \in \{\emptyset,[p,t,k]\}u\{\emptyset,[p,t,k]\}$. Pattern A has two place restrictions on the coda; Pattern B has one on the coda and one on the onset:

	Pattern A	Pattern B
Unviolated	*[–syll, +Dor]]$_\sigma$ (=NoDorCoda)	*[–syll, +Dor]]$_\sigma$ (=NoDorCoda)
constraints	*[–syll, +Lab]]$_\sigma$ (=NoLabCoda)	*$_\sigma$[[–syll, +Lab] (=NoLabOns)
Positive	u, ut, pu, put, tu, tut, ku, kut	u, up, ut, tu, tup, tut, ku, kup, kut
Negative	up, uk, pup, puk, tup, tuk, kup, kuk	uk, pu, pup, put, tuk, kuk

(27) To make analysis easier, turn off recombination to make all reproduction asexual mutation. The mutation distances between the critical constraints are then

a. *Condition A*: 2. From *[–syll, +Lab]]$_\sigma$ to *[–syll, +Dor]]$_\sigma$, delete [+Lab], insert [+Dor]
b. *Condition B*: 4. From *[–syll, +Lab]]$_\sigma$ to *$_\sigma$[[–syll, +Dor], delete [+Lab], insert [+Dor], unset right boundary, set left boundary.

(Same holds for other micro-constraints that instantiate these macro-constraints)

(28) Discovering either of the critical constraints should therefore prime discovery of the other better in the A condition than in the B condition. Concretely, we expect that in Condition A, as compared to Condition B,

a. time between discovery of the two constraints will be smaller
b. the weights of the two constraints will be more strongly correlated
c. the harmonies of the critical stimuli (labial- and dorsal-final in Condition A, labial-initial and dorsal-final in Condition B) will be more strongly correlated

(29) Simulation parameters: Mutation rate of $\mu = 0.01$[1] learning rate of $\epsilon = 0.05$, a population of $N = 1000$ constraints initialized to `*(L PrWd R)`, and a weight quantum of $\zeta = 0.01$. Time limit of 500 errors, 1000 trials, or 24 hours of real time. 20 paired replications in each condition.

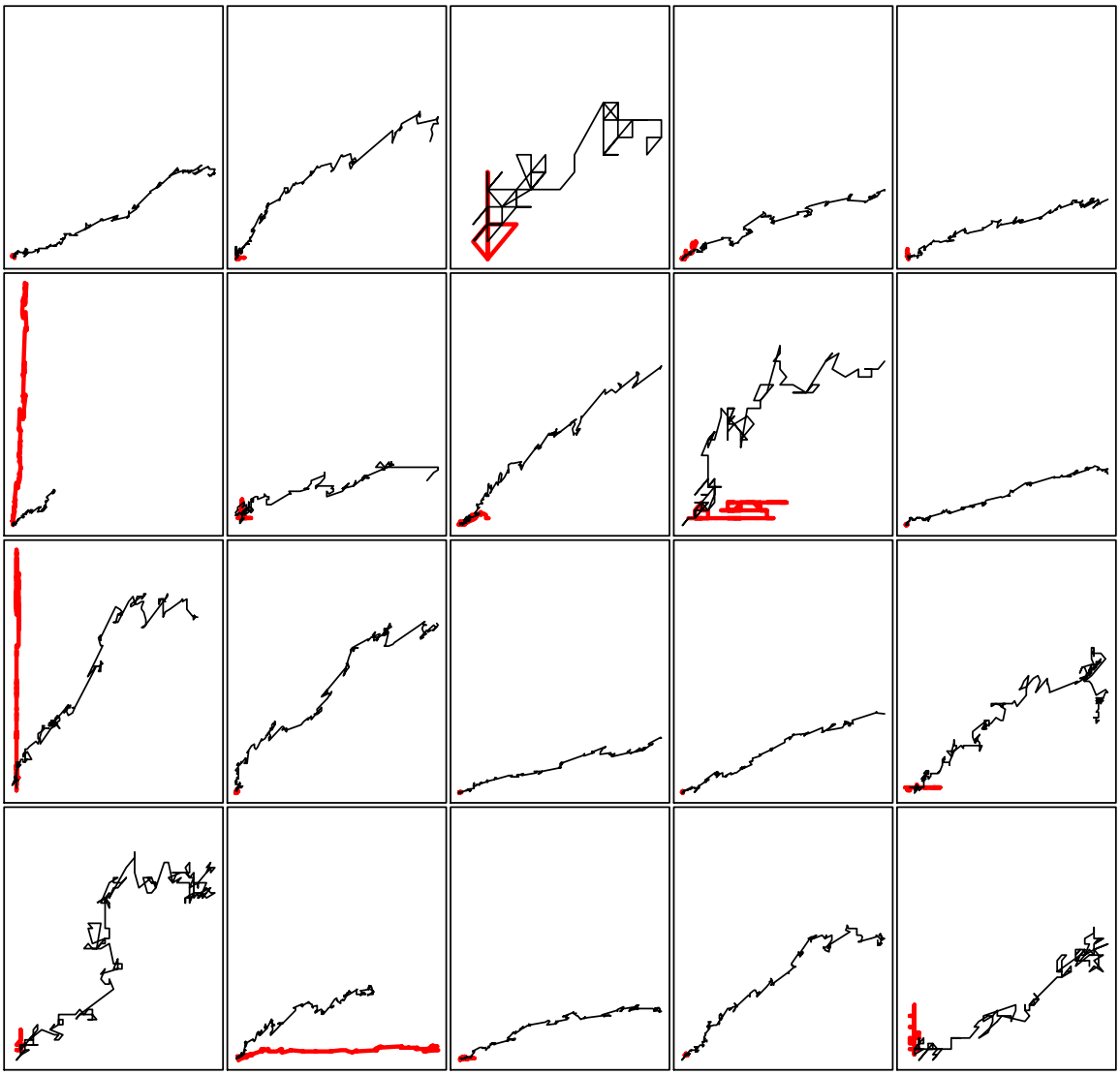
(30) *Prediction: Time between discovery smaller in A than B*: ✓

		Discovery of		
		*[–syll, +Dor]]$_\sigma$	*$_\sigma$[[–syll, +Lab] *or* *[–syll, +Lab]]$_\sigma$	Abs. diff.
Median:	A	144	168	48
	B	433	335	271

The difference was smaller for the A condition than the B condition in 13 pairs, longer in 3 pairs, and NA in 4 pairs (the ones where the B learner never discovered either constraint). The difference is significant (`binom.test (13,3)`: 95% CI is [0.54, 0.96], $p = 0.021$).

(31) *Prediction: Weights of the two constraints more strongly correlated in A than B*: ✓ Horizontal axis: w(*[–syll, +Dor]]$_\sigma$) in both conditions. Vertical axis: w(*[–syll, +Lab]]$_\sigma$) in A (black), w(*$_\sigma$[[–syll, +Lab]) in B (red).

[1] I.e., the probabilities of the following mutations were set to 0.01: add prosodic constituent, delete prosodic constituent, toggle prosodic boundary assertion, gain unary feature, lose unary feature, gain binary feature, lose binary feature, invert binary-feature coefficient.



(32) *Prediction: Harmonies of the critical stimuli more strongly correlated in A than B*: ✓ (Plots omitted; look almost exactly like those in (31).

(33) *Attention-like effects*: Clues in the data can cause the learner to search some regions of constraint space more intensively. Here, the constraint *[–syll, +Dor]]$_\sigma$ (critical in A and B conditions) is one mutation away from *[–syll]]$_\sigma$ (i.e., NoCoda). As the table shows,

	Median trial number			
Event	A		B	
*[–syll]]$_\sigma$ discovered	20		21	
*[–syll]]$_\sigma$ weight peaks (value)	591	(0.865)	237	(0.1)
*[–syll, +Dor]]$_\sigma$ discovered	144		443	

a. *[–syll]]$_\sigma$ is discovered early in both A and B

b. *[–syll]]$_\sigma$ is better supported in A (4/8 pos. vs. 0/8 neg.) than in B (3/8 pos. vs. 1/8 neg.),so once discovered, it gains weight faster:

	peak weight/time to peak
A	$0.865/591 = 0.00146$
B	$0.1/237 = 0.00042$

c. More weight in *[–syll]]$_\sigma$ means more opportunities to spawn *[–syll, +Dor]]$_\sigma$, so A gets there first.

d. The *A* learner "notices" that codas matter, i.e. up-weights *[–syll]]$_\sigma$.

e. That "focuses its attention" on the coda position (by allowing the approximate solution *[–syll]]$_\sigma$ to elbow out other constraints).

f. This "focused attention" results in a more-intensive search among neighbors of *[–syll]]$_\sigma$, which soon finds *both* critical constraints. R&D work that helps find one constraint also helps find the other.

(The critical constraints then outcompete the approximate constraint and drive its weight down.)

g. In the *B* condition, it takes longer to discover the critical constraint because "attention" is divided between the onset and coda positions (the data does not "call attention" to one more than the other).

---

7 Discussion

(34) *Inductive biases*: Constraint interaction via Max Ent HG + weight update via fitness-based selection + constraint generation via mutation and recombination → characteristic inductive biases which affect both the end state and the learning path. How might they show up empirically, in the lab or in nature?

a. *Sharing: The learner favors grammars where the macro-constraints share formal components.* (E.g., NoDorsalCoda and NoLabialCoda in §6.)

Using data sampled from P-Base (Mielke, 2008), Carter (2017) found that languages in fact tend to re-use phonological features: The probability that a language which uses Feature *F* in *N* classes uses it in $N+1$ classes increases with *N* (i.e., a preferential attachment process).

Sharing bias is distinct from a generality bias towards patterns that are supported by multiple overlapping constraints (Pater and Moreton, 2012; Moreton, 2012).

b. *Oversharing: Intermediate grammars can overgeneralize the shared components.* (E.g., NoCoda in §6.)

In segment-class-learning experiments with adults, preference for conforming-old and conforming-new over nonconforming segments precedes preference for conforming-old over conforming-new segments (Linzen and Gallagher, 2014, 2017). The Hayes and Wilson (2008) Max Ent learner predicts the opposite time course (Linzen and O'Donnell, 2015).

c. *Nepotism: Weighty macro-constraints can maintain relatives above their rightful level (even in the end state).* (E.g., NoCoda is continually replenished from NoDorsalCoda and NoLabialCoda as the simulation continues.)

In segment-class learning with adults, generalization to untrained segments is stronger when they are more similar to trained segments (Cristiá et al., 2013). Prickett (2018) showed that GMECCS (Moreton et al., 2017) underpredicts the difference, but that the fit can be improved by making weight updates "leak" between featurally-similar constraints.

The predictions here may be bizarre; e.g., that a language with high-weighted NoCoda would also show TETU effects against onsets, since NoOnset is a close relative.

If the mutation rate decreases later in learning, nepotism is reduced.

(35) *Limiting behavior*: Varying the parameters causes Evolutionary Winnow-Max Ent to resemble qualitatively different models:

a. *Large population, small weight quantum*: Approximates a constraint-based model whose initial constraint set contains all possible constraints up to a certain size (IMECCS/GMECCS, Pater and Moreton 2012; Moreton et al. 2017). The reason is that the mutants created on any error trial will sample the space of possible constraints densely.

b. *Small population, large weight quantum*: Approximates a serial hypothesis-tester that keeps trying categorical rules until it finds one that works. These models, common in the psychology literature on concept learning, normally incorporate a bias towards syntactically "simple" rules (Shepard et al., 1961; Nosofsky et al., 1994; Feldman, 2006; Ashby et al., 2011; Goodwin and Johnson-Laird, 2013). Can be achieved here by setting initial state to a simple constraint, and making deletion more likely than insertion in mutation.

c. *Stimuli (candidates) added to constraint set* after each trial: Approximates a learning strategy of memorizing stimuli. Mutation causes gradual "forgetting".

⇒ May provide continuity with psychological models of non-linguistic category learning (Feldman, 2006; Vigo, 2013); Gluck and Bower 1988; Shepard et al. 1961; Nosofsky et al. 1994; Feldman 2000; Mathy and Bradmetz 2004; Feldman 2006; Lafond et al. 2007; Bradmetz and Mathy 2008; Vigo 2009; Goodwin and Johnson-Laird 2011; Kurtz et al. 2013; Gluck and Bower 1988; Moreton et al. 2017.

References

Adriaans, F. and R. Kager (2010). Adding generalization to statistical learning: the induction of phonotactics from continuous speech. *Journal of Memory and Language 62*(3), 311–331.

Alderete, J. (1999). *Morphologically governed accent in Optimality Theory*. Ph. D. thesis, University of Massachusetts, Amherst.

Anderson, S. R. (1981). Why phonology isn't "natural". *Linguistic Inquiry 12*, 493–539.

Ashby, F. G., L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldron (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review 105*(3), 442–481.

Ashby, F. G., E. J. Paul, and W. T. Maddox (2011). COVIS. In E. M. Pothos and A. J. Willis (Eds.), *Formal approaches in categorization*, Chapter 4, pp. 65–87. Cambridge, England: Cambridge University Press.

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.

Becker, M. (2009). *Phonological trends in the lexicon: the role of constraints*. Ph. D. thesis, University of Massachusetts, Amherst.

Becker, M. and A. Tessier (2011). Trajectories of faithfulness in child-specific phonology. *Phonology 28*, 163–196.

Benua, L. (1997). *Transderivational identity: phonological relations between words*. Ph. D. thesis, University of Massachusetts, Amherst, Mass.

Berent, I., C. Wilson, G. F. Marcus, and D. K. Bemis (2012). On the role of variables in phonology: Remarks on hayes and wilson 2008. *Linguistic inquiry 43*(1), 97–119.

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

Bradmetz, J. and F. Mathy (2008). Response times seen as decompression times in boolean concept use. *Psychological Research 72*(2), 211–234.

Buckley, E. (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages*, Volume 9 of *UCSB Working Papers in Linguistics*.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review 67*(6), 380–400.

Carter, W. T. (2017). Phonological activeness effects in language acquisition and language structuring. Senior Honors thesis, Department of Linguistics, University of North Carolina, Chapel Hill.

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.

Coetzee, A. W. and J. Pater (2008). Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language & Linguistic Theory 26*(2), 289–337.

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., J. Mielke, R. Daland, and S. Peperkamp (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology 4*, 259–285.

Dietrich, A. and H. Haider (2015). Human creativity, evolutionary algorithms, and predictive representations: the mechanics of thought trials. *Psychonomic Bulletin and Review 22*, 897–915.

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

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

Flack, K. (2007). Templatic morphology and indexed markedness constraints. *Linguistic Inquiry 38*(4), 749–758.

Fukazawa, H. (1999). *Theoretical implications of OCP effects on features in Optimality Theory*. Ph. D. thesis, University of Maryland, College Park.

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

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.

Goodwin, G. P. and P. N. Johnson-Laird (2011). Mental models of Boolean concepts. *Cognitive Psychology 63*(34–59).

Goodwin, G. P. and P. N. Johnson-Laird (2013). The acquisition of Boolean concepts. *Trends in Cognitive Sciences 17*(3), 128–133.

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

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.

Inkelas, S. (2008). The morphology-phonology connection. In *Proceedings of the Berkeley Linguistics Society*, Volume 34, Berkeley, California, pp. 145–162. Berkeley Linguistics Society and Linguistic Society of America.

Ito, J. and A. Mester (2001). Covert generalizations in Optimality Theory: the role of stratal faithfulness constraints. *Studies in Phonetics, Phonology, and Morphology 7*, 3–33.

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.

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.
Kurtz, K. J., K. R. Levering, R. D. Stanton, J. Romero, and S. N. Morris (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*(2), 552–572.
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.
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.
Linzen, T. and G. Gallagher (2014). The timecourse of generalization in phonotactic learning. In *Proceedings of the Annual Meetings on Phonology*, Volume 1.
Linzen, T. and G. Gallagher (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology 8*(1).
Linzen, T. and T. J. O'Donnell (2015, September 17–21). A model of rapid phonotactic generalization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, pp. 1126–1131. Association for Computational Linguistics.
Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning 2*, 285–318.
Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review 9*(4), 829–835.
Luce, R. D. (2005 [1959]). *Individual choice behavior: a theoretical analysis*. New York: Dover.
Maddox, W. T. and F. G. Ashby (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes 66*, 309–332.
Magri, G. (2013). HG has no computational advantages over OT: toward a new toolkit for computational OT. *Linguistic Inquiry 44*(4), 569–609.
Marcus, G. F., S. Vijayan, S. B. Rao, and P. M. Vishton (1999). Rule learning by seven-month-old infants. *Science 283*, 77–80.
Mathy, F. and J. Bradmetz (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition/Cahiers de Psychologie Cognitive 22*(1), 41–82.
McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry 12*, 373–418.
McCarthy, J. J. and A. M. Prince (1993). Generalized alignment. In G. Booij and J. van Marle (Eds.), *Yearbook of morphology 1993*, pp. 79–153. Kluwer.
Mielke, J. (2008). *The emergence of distinctive features*. Oxford, England: Oxford University Press.
Minsky, M. and S. Papert (1969). *Perceptrons: an introduction to computational geometry*. Cambridge, Massachusetts: MIT Press.
Moreton, E. (2010a, April). Connecting paradigmatic and syntagmatic simplicity bias in phonotactic learning. Department colloquium, Department of Linguistics, MIT.
Moreton, E. (2010b, February). Constraint induction and simplicity bias. Talk given at the Workshop on Computational Modelling of Sound Pattern Acquisition, University of Alberta.
Moreton, E. (2010c, May). Constraint induction and simplicity bias in phonotactic learning. Handout from a talk at the Workshop on Grammar Induction, Cornell University.
Moreton, E. (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language 67*(1), 165–183.
Moreton, E. (2018). Conditions on abruptness in a gradient-ascent Maximum Entropy learner. In G. Jarosz and J. Pater (Eds.), *Proceedings of the Society for Computation in Linguistics*, Volume 1, pp. Article 13.
Moreton, E. (2019). Constraint breeding during on-line incremental learning. In *Proceedings of the Society for Computation in Linguistics*, Volume 2, pp. Article 9.
Moreton, E., J. Pater, and K. Pertsova (2017). Phonological concept learning. *Cognitive Science 41*(1), 4–69.
Nakano, K., H. Hiraki, and S. Ikeda (1995). A learning machine that evolves. In *Proceedings of ICEC–95*,

pp. 808–813.
Nosofsky, R. M., T. J. Palmeri, and S. C. McKinley (1994). Rule-plus-exception model of classification learning. *Psychological Review 101*(1), 53–79.
Ota, M. (2004). The learnability of the stratified phonological lexicon. *Journal of Japanese Linguistics 20*, 19–40.
Pater, J. (2000). Non-uniformity in English secondary stress: the role of ranked and lexically specific constraints. *Phonology 17*, 237–274.
Pater, J. (2007). The locus of exceptionality: morpheme-specific phonology as constraint indexation. In L. Bateman, M. O'Keefe, E. Reilly, and A. Werle (Eds.), *Papers in Optimality Theory III*, pp. 259–296. Amherst: Graduate Linguistics Students Association, University of Massachusetts.
Pater, J. (2009). Morpheme-specific phonology: constraint indexation and inconsistency resolution. In S. Parker (Ed.), *Phonological argumentation: essays on evidence and motivation*, pp. 1–33. London: Equinox.
Pater, J. (2014). Canadian Raising with language-specific weighted constraints. *Language 90*(1), 230–240.
Pater, J. and E. Moreton (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad 3*(2), 1–44.
Pizzo, P. (2013, January 19). Learning phonological alternations with online constraint induction. Slides from a presentation at the 10th Old World Conference on Phonology (OCP 10).
Prickett, B. (2018). Similarity-based phonological generalization. In G. Jarosz and J. Pater (Eds.), *Proceedings of the Society for Computation in Linguistics*, Volume 1, pp. Article 24.
Prince, A. and P. Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Department of Linguistics, Rutgers University.
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.
Shepard, R. N., C. L. Hovland, and H. M. Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs 75*(13, Whole No. 517).
Simonton, D. K. (1999). Creativity as blind variation and selective retention: is the creative process Darwinian? *Psychological Inquiry 10*(4), 309–328.
Simonton, D. K. (2004). *Creativity in science: chance, logic, genius, and Zeitgeist*. Cambridge University Press.
Smith, J. D., M. E. Berg, R. G. Cook, M. S. Murphy, M. J. Crossley, J. Boomer, B. Spiering, M. J. Beran, B. A. Church, F. G. Ashby, and R. C. Grace (2012). Implicit and explicit categorization: a tale of four species. *Neuroscience and Biobehavioral Reviews 36*(10), 2355–2369.
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.
Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology 50*, 203–221.
Vigo, R. (2013). The GIST of concepts. *Cognition 129*, 138–162.
Wilson, C. and G. Gallagher (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry 49*(3), 610–623.