Evolving constraints and rules in Harmonic Grammar

E. Moreton

1 Introduction

(1) Research program: *Comparative study of inductive learning.* What has linguistic learning got to do with human inductive learning in other domains?

(2) *In particular:* What has phonotactic learning got to do with inductive learning of non-linguistic concepts? Analogous inductive problems can arise across domains; are they solved in analogous ways?

a. *Phonology*

	Consonant	
Vowel	short	long
short	*lam	lamm
long	laːm	*laːmm

Swedish: Either the vowel or the consonant of a closed stressed syllable is long, but not both (Löfstedt, 1992).

b. *Morphology*

	Number	
Case	sing.	pl.
Acc.	mur	mur-s
Nom.	mur-s	mur

Old French: /-s/ is attached to an *o*-stem noun if it is nominative or plural, but not both (Luquiens, 1909, §289).

c. *Non-linguistic concept*

	Adaptive immune system	
Backbone	Absent	Present
Present	(none)	Vertebrates
Absent	Invertebrates	(none)

Non-linguistic concept: An animal species has a backbone if and only if it has an adaptive immune system (Litman et al., 2010).

(3) Today's focus: *Gradual weight update vs. serial hypothesis testing.* Studies of non-linguistic (mainly visual) concept learning have led psychologists to hypothesize two distinct learning processes that have different properties and that are facilitated by different experimental conditions (Ashby et al. 1998; Love 2002; Maddox and Ashby 2004; Smith et al. 2012; also with language-like stimuli, Endress and Bonatti 2007; Endress and Mehler 2009; Weinert 2009).

Explicit system (≈ reasoning)	*Implicit system* (≈ intuition)
Effortful	Effortless
Conscious	Unconscious
Abrupt	Gradual
Demands attention and working memory	Does not need attention or working memory
Learns Type IFF/XOR ("Type II") patterns faster than family-resemblance ("Type IV") patterns	Learns Type IV patterns faster than Type II patterns
Use is facilitated by supervised training, instructions to seek a rule, verbalizable features	Use is facilitated by unsupervised training, instructions that don't mention rules, non-verbalizable features
Can be modeled as serial testing of featurally-simple hypotheses ("serial hypothesis testing")	Can be modeled as weight update on array of property detectors ("gradual weight update")

(4) The proposed signatures of both learning modes can be found in phonological experiments (Moreton and Pertsova, 2016).

[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 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.

(5) $\boxed{\text{Idea for today's talk:}}$ *What unites the two kinds of learning is that both depend on finding formulas (categorical, symbolic objects) by trial and error.*

This talk describes an evolutionary algorithm for simultaneously inducing, weighting, and applying Harmonic Grammar constraints, and shows how it can approximate both gradual weight update and serial hypothesis-testing.

(6) Talk map:

§2 Brief précis of the Evolutionary Winnow-MaxEnt Subtree learner

§3 How it approximates gradual weight update, illustrated using an unsupervised phonological learning experiment

§4 How it approximates serial hypothesis-testing, illustrated using a supervised visual learning experiment

§5 Discussion

---

2 The Evolutionary Winnow-MaxEnt-Subtree Learner

(7) Précis of the Evolutionary Winnow-MaxEnt Subtree learner (Moreton 2010a,b,c, 2019, 2020; URL for code and replication kit in Moreton 2020):

a. Candidates are trees. Markedness constraints are subtrees. (No faithfulness, yet.)

b. Constraint weights are population sizes.

c. Weight update is reproduction, inducing a nondeterministic variant of Winnow-2 (Littlestone, 1988; Moreton, 2019)

d. Constraint induction is evolution (variation and differential reproductive success)

(8) Why evolution?

a. Is an established technology for efficiently searching large, inconveniently-shaped hypothesis spaces (Bäck, 1996; Eiben and Smith, 2003; De Jong, 2006)

b. Hasn't been tried in phonology yet, though it has been applied successfully to related problems such as evolving receptive fields for inputs to the single-layer perceptron (Nakano et al., 1995) and evolving tree structures (Cramer, 1985; Koza, 1989).

c. Allows constraints to be induced and weighted simultaneously, and on-line rather than in batch mode; hence promising as account of what humans do

d. Connects phonological learning with a leading theory of human creativity in other domains (Campbell, 1960; Simonton, 1999, 2004; Kronfeldner, 2010; Dietrich and Haider, 2015).

e. Connects gradual-reweighting models with serial hypothesis-testing models. (*This talk*)

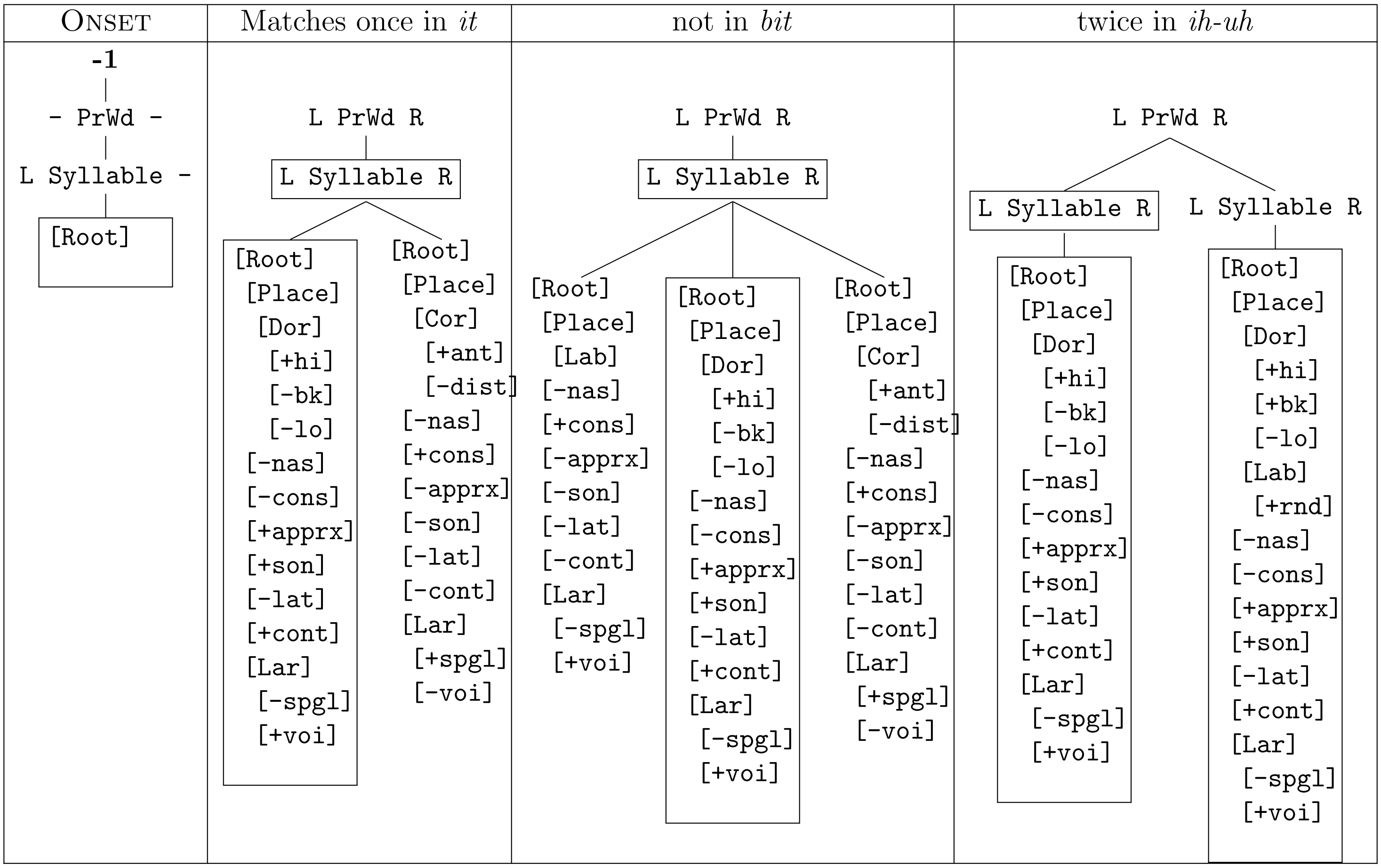
2.1 Constraints and candidates

(9) Constraints and candidates are consubstantial (Golston 1996; Burzio 1999; see also Futrell et al. 2017):

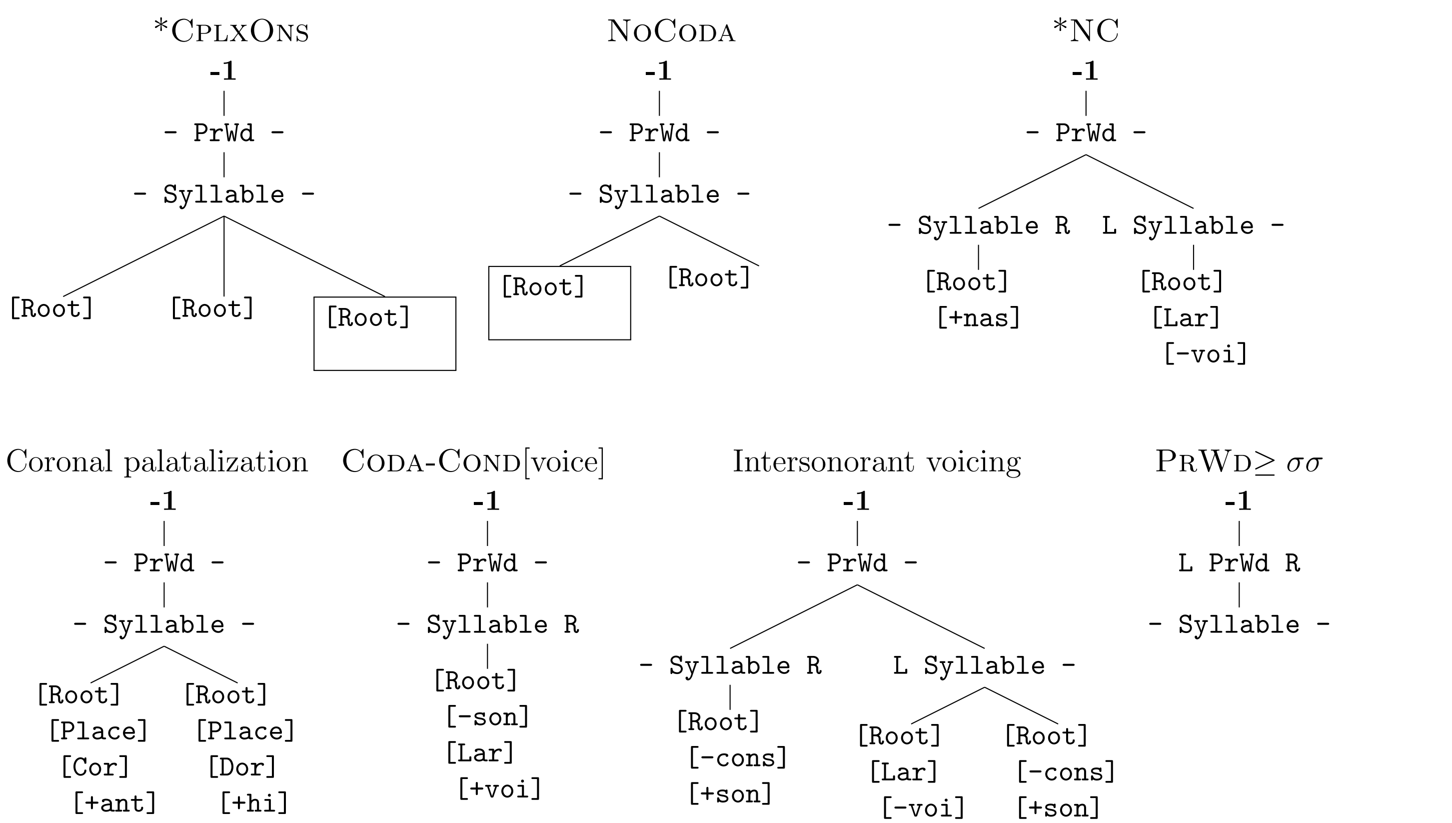
a. *Candidates are trees* using standard Feature-Geometric tree structure (Goldsmith, 1976; McCarthy, 1981; Sagey, 1990; Clements and Hume, 1995). This implementation uses a slightly simplified version of the one in Gussenhoven and Jacobs (2005, Ch. 5).

b. *Constraints are subtrees*: A constraint is a (possibly incomplete) representation which describes a locus of violation or of satisfaction, plus an associated number of marks.

(10) This example illustrates Onset, à la Smith (2006), as applied to example candidates. (The box marks the head of a prosodic category.)



(11) This scheme is flexible enough to express a wide range of constraints:



(12) Adding Greek-letter variables (not discussed in this talk; see Moreton 2010c) allows the schema to express assimilation and dissimilation:

```
        *AGREE-[high]            Nasal Place Assimilation          Syllable alliteration
              -1                            -1                              -1
              |                             |                               |
          - PrWd -                      - PrWd -                        - PrWd -
         /        \                    /        \                      /        \
- Syllable -   - Syllable -   - Syllable R   L Syllable -   L Syllable -   L Syllable -
      |              |               |              |              |              |
 [Root]          [Root]         [Root]          [Root]         [−α Root]      [α Root]
  [Place]         [Place]         [α Place]       [α Place]
   [Dor]           [Dor]          [+nas]          [-cont]
    [+αhi]          [-αhi]
```

(13) Properties of the Subtree Schema:

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

b. Supports both adjacent and non-adjacent dependencies (e.g., Nasal Place Assimilation and Agree-[high] in (12))

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

d. Lends itself to recursive recombination and mutation (see below)

2.2 Micro- and macro-constraints

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

Macro-constraints:	*CpOns			Max			
Weights:	4			3			
Micro-constraints:	*CpOns	...	*CpOns	Max	...	Max	
Weights:	0.01	*398 more*	0.01	0.01	*298 more*	0.01	
/bʃib-dʒu/							
[bʃib.dʒu]	*	...	*		...		$H = -4$
→[ʃib.dʒu]		...		*	...	*	$H = -3$

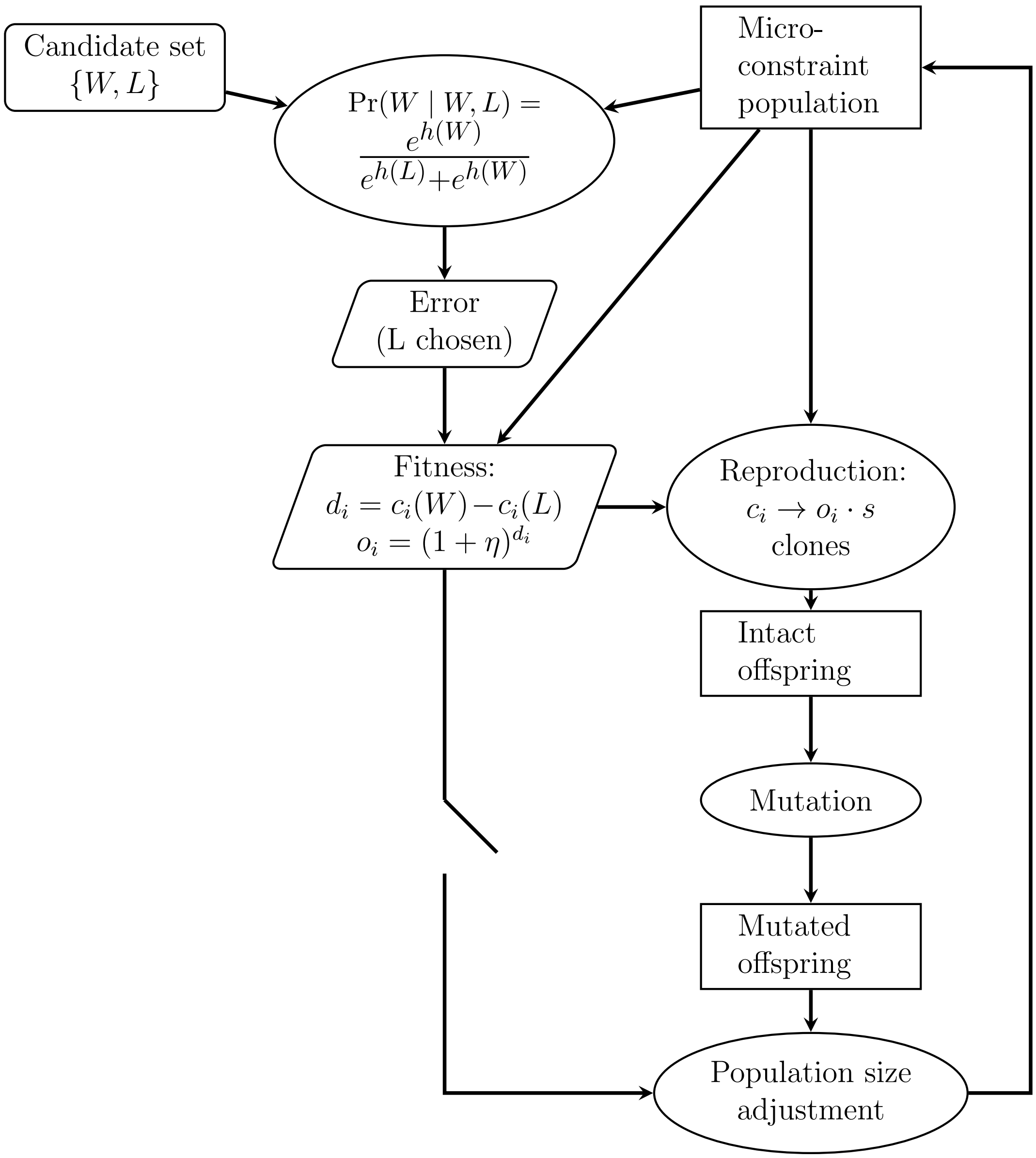
(15) *Macro-constraints are equivalence classes*: The algorithm itself sees only micro-constraints. For analytic convenience, we can define two micro-constraints as belonging to the same macro-constraint if they assign the same scores to all candidates (i.e., if they are notational variants of each other).

2.3 Evolving constraints

(16) *Weight update is reproduction*, inducing a nondeterministic variant of the Winnow-2 algorithm (Littlestone, 1988; Moreton, 2019).

(17) *Constraint induction is evolution*, i.e., reproduction with variation and selection, i.e., trial and error.

(18) A map of the algorithm:



Several other features (recombination, meta-constraints, candidate memorization, momentum, etc.) are not shown because they were turned off for the simulations discussed here; see paper for details.

3 Approximation to gradual reweighting

(19) Large population size N plus small weight quantum ζ plus fitness-insensitive population-size adjustment means that

a. macro-constraints approximate HG constraints with continuous weights

b. mutants created on any error sample the space of possible micro- constraints densely

c. model's only record of past success of macro-constraints is their population size

⇒ learner should approximate a model with continuous, gradually-updated weights and a rich pre-specified constraint set (e.g., the Configural Cue Model of Gluck and Bower 1988a,b, the Gradual Learning Algorithm of Boersma 1997; Boersma and Hayes 2001, or the IMECCS/GMECCS model of Pater and Moreton 2012; Moreton et al. 2017).

(20) How could we tell if it's doing it right? A characteristic of such models is that they learn single-feature ("Type I") patterns faster than three-way gang-effect ("Type IV") patterns, and those faster than exclusive-or ("Type II") patterns (Moreton et al., 2017). Real-life examples of the relevant patterns, located by analyzing P-Base (Mielke, 2008; Moreton and Pertsova, 2014).

	[–back]		[+back]	
	[–rnd]	[+rnd]	[–rnd]	[+rnd]
[+high]	$\boxed{\text{i}}$	$\boxed{\text{y}}$	ɨ	u
[–high]	$\boxed{\text{e}}$	$\boxed{\text{ø}}$	a	o

Type I: The vowel inventory of Turkish. Boxes enclose vowels which cause secondary palatalization of adjacent /k g/ (Bateman, 2007, 71).

	[–voice]		[+voice]	
	[–distr]	[+distr]	[–distr]	[+distr]
[–cont]	t	p, tʃ, k	$\boxed{\text{n}}$	$\boxed{\text{m}}$
[+cont]	$\boxed{\text{s}}$	$\boxed{\text{ʃ, x, h}}$	l	w,j

Type II: The consonant inventory of Unami Delaware (Goddard, 1979). Boxes enclose sounds that can precede non-coronal stops; they are [+cont] iff [−voice].

	[–back]		[+back]	
	[–rnd]	[+rnd]	[–rnd]	[+rnd]
[+high]	$\boxed{\text{i}}$	$\boxed{\text{y}}$	$\boxed{\text{ɨ}}$	u
[–high]	$\boxed{\text{e}}$	ø	a	o

Type IV: The boxes enclose those Kirghiz vowels which undergo raising and tensing before palatal consonants (Hebert and Poppe, 1963, 3–7). I.e., "anything within one feature of /i/".

(21) Humans doing unsupervised phonological learning exhibit the same difficulty order. Illustration: Moreton et al. (2017, Exp. 1).

a. Stimuli: MBROLA-synthesized $C_1V_1C_2V_2$ words with inventory /t k d g/ /i u æ ɔ/ (Moreton, 2008; Lin, 2009; Kapatsinski, 2011; Moreton, 2012).

b. Phonotactic patterns: For each participant, 3 of the 8 stimulus features were randomly chosen as the relevant features, and then randomly mapped onto the three logical features defining the Shepard pattern to produce the "language" for that participant. Examples:

L1 (TYPE I):	C1 is voiceless **t**igu, **k**ada, **t**ika, **k**ugu, . . .
L2 (TYPE II):	C1 is voiced iff C2 is voiceless. **d**i**k**u, **t**æ**g**i, **k**a**g**æ **g**a**t**a, . . .
L3 (TYPE IV):	At least two of: C1 is voiced, C2 is dorsal, V2 is back **k**a**ga**, **g**a**gu**, **g**æ**ku**, **t**æ**ki**, . . .

c. *Instructions*: Participants (who were run in a lab, by a human) were told they would learn to pronounce words in an artificial language, and then be tested on ability to recognize words in that language.

d. *Training*: They listened to and repeated aloud 32 randomly-chosen pattern-conforming stimuli 4 times over.

e. *Test*: Then they heard 32 randomly-chosen pairs of new stimuli (one pattern-conforming, one not) and tried to identify the one that was "a word in the language you were studying".

f. *Results*: $I > IV > II$ order, matched by GMECCS

(22) Can the Evolutionary Winnow-MaxEnt Subtree learner replicate this difficulty order? *Simulation 2* from Moreton (2020):

a. $N = 2000$ micro-constraints, $\zeta = 0.05$ weight units per micro-constraint. Population adjustment step ignored fitness. 100 replications in each condition.

b. Trained using same phonological stimuli as humans (same 32 words, repeated 4 times). Losers were generated by sampling from the distribution specified by the current grammar (Jäger, 2007).

c. Tested on same phonological stimuli as humans (same 32 novel pairs)

(23) Proportion pattern-conforming responses in the test phase ($\pm$ 1 s.d., not s.e.m.) for Simulation 2, GMECCS (Moreton et al., 2017, Figure 10), and human data (Moreton et al., 2017, Table 5), showing $I > IV > II$ order.

	Pattern type		
	I	II	IV
Simulation 2	0.83 ± 0.13	0.48 ± 0.02	0.60 ± 0.05
GMECCS	0.72	0.58	0.66
Human	0.73 ± 0.12	0.57 ± 0.11	0.70 ± 0.09

$\Rightarrow$ As expected, a large population size and a small weight quantum yielded the $I > IV > II$ performance characteristic of a gradient-ascent Max-Ent learner with a rich set of prespecified constraints.

(24) How did it happen?

a. Many simulated participants found a wholly valid constraint for Type I, or a partially-valid single-feature approximation for Type IV. But Valid constraints for Type II were hard to find — *no* simulated participant in this run found even one.

b. That happened because of the large mutation distance between the initial constraints (`*[+wug]`, `*[-wug]`) and the deep forked tree required by the Subtree Schema for the two-segment Type II pattern.

c. Different from how $I > IV > II$ arises in GMECCS, where all necessary constraints were furnished in advance and Type II was slow because the positive stimuli had fewer positive neighbors than in Types I and IV (Moreton et al., 2017, §4.1.2).

4 Approximation to serial hypothesis testing

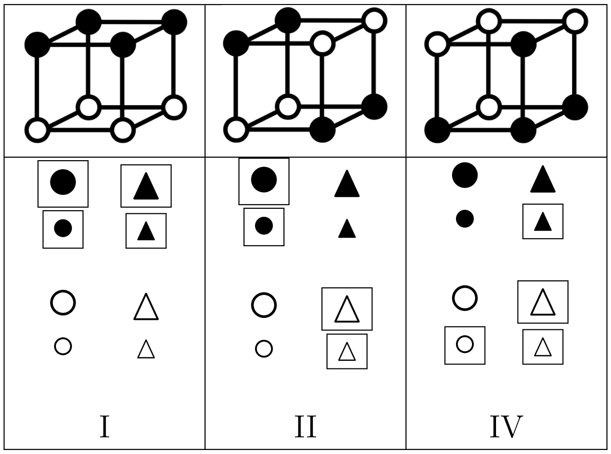
(25) Small population size N plus large weight quantum ζ plus fitness-based population-size update means

a. each micro-constraint affects a candidate's harmony so much as to be in effect a categorical rule

b. making one error can ban a micro-constraint permanently (error $\rightarrow$ low fitness $\rightarrow$ eliminated; when re-innovated, it starts off with the same low fitness and so doesn't make the cut at the population-adjustment step)

⇒ should approximate a serial hypothesis-tester that keeps trying one categorical rule after another until it finds one that works (e.g., Nosofsky et al. 1994; Feldman 2006; Ashby et al. 2011).

(26) The characteristic order of difficulty in visual concept-learning is $I > II > IV$ (e.g., Shepard et al. 1961; Nosofsky et al. 1994; Smith et al. 2004; see critical review in Kurtz et al. 2013).



Pattern Types I, II, and IV of Shepard et al. (1961), illustrated using visual stimuli. Type I is defined by **one** feature ("the figure is black"); Type II is an iff/xor relation between **two** features ("black iff round"); and Type IV is a **three**-feature gang effect ("at least two of white, triangular, small").

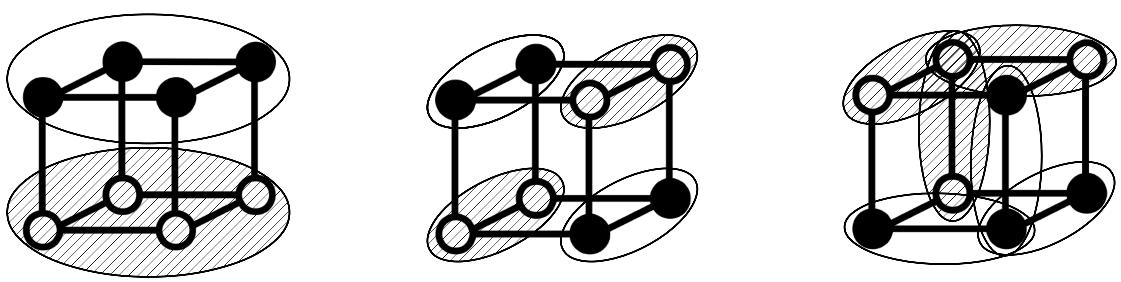
(27) Serial hypothesis-testing models in the concept-learning literature account for $I > II > IV$ by positing a hard-wired bias towards hypotheses which involve fewer features (Shepard et al., 1961; Nosofsky et al., 1994; Feldman, 2006; Ashby et al., 2011; Goodwin and Johnson-Laird, 2013).

(28) The bias in the Winnow-MaxEnt-Subtree learner has a different origin.

a. For small population size N and large weight quantum ζ, the handful of constraints act like individual categorical prohibitions.

b. The minimal large-ζ grammar for each Type is shown below. E.g., Type I requires only constraints about the color feature (vertical axis): `*[-wug, +black]` (clear oval) and `*[+wug, -black]` (hatched oval).

I | II | IV



2 constraints | 4 constraints | 6 constraints

c. ⇒ Small N/large ζ should favor Type I over Type II, and Type II over Type IV.

(29) Simulation 3 used $N = 7, \zeta = 12$.

a. High mutation rate plus large clutch size ($s = 12$) made the offspring population a diverse random sample from the 54 possible constraints.

b. Fitness-based population adjustment plus fitness memory (a re-innovated lost constraint resumes the fitness it last had) gradually eliminates constraints that favor losers

c. Offspring population becomes more and more a random sample of size 7 from the valid constraints

d. A random sample of size 7 valid constraints is more likely to solve Type I than Type II, and Type II than Type IV (see paper for details).

(30) Attainment of criterion performance (32 consecutive correct responses in 400 trials) for Simulation 3 and human participants (Nosofsky et al., 1994, 356). Mean trials to criterion excludes cases where criterion was not reached. There were 100 replications.

	% participants reaching criterion			Mean trials to criterion		
	I	II	IV	I	II	IV
Sim.	100	98	74	68	161	210
Human	100	100	100	44	85	127

(31) ⇒ Changing the model parameters has caused Types II and IV to change places with respect to Simulation 2. The order of difficulty, $I > II > IV$, is the same for the learner as for the humans (who are about 40% faster in all conditions).

Smaller values of N amplified the advantage of Type II over Type IV. For $N \leq 5$, *no* Type IV simulations reached criterion.

(32) Check: Is the difference between Simulations 2 and 3 really due to the change in learning parameters (and not to the change in the stimulus space)? Yes: When the parameters are set as in Simulation 2 (large population, small weight quantum, fitness-insensitive population adjustment), the $I > IV > II$ order is restored:

	% participants reaching criterion			Mean trials to criterion		
	I	II	IV	I	II	IV
Sim.	100	100	100	70	196	155

5 Discussion

(33) *Big Question:* How is phonological learning like or unlike learning in other domains?

(34) Proposed phonological learning models have been architecturally very different from proposed models for visual pattern learning: gradual weight update vs. serial hypothesis testing.

(35) But both types of model have something in common: *They both depend on finding formulas (symbolic objects) by trial and error.*

a. Obvious for serial hypothesis testers.
b. Phonological learners that use gradual weight update *must* induce at least some of their constraints from the phonological data (e.g., constraints that are specific to particular lexemes, lexical strata, inflectional paradigms, etc., or that refer to phonetically-arbitrary or otherwise idosyncratic patterns — see refs in paper)

(36) In the Evolutionary Winnow-MaxEnt-Subtree model, we have a single model whose behavior can approximate that of two other model types.

a. Large population/small weight quantum ≈ parallel search for constraints
b. Small population/large weight quantum ≈ serial search for categorical rules

(37) Specifically, the learner can get both of two signature difficulty orders for the Shepard et al. 1961 pattern types:

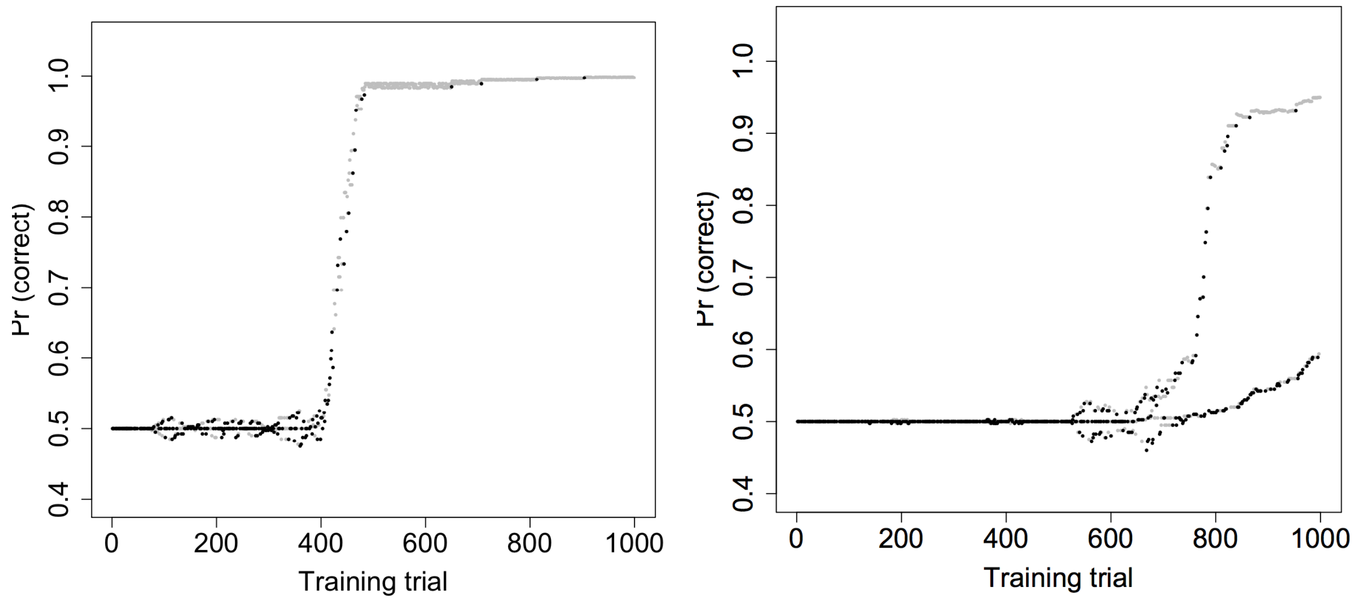
a. $I > IV > II$:
   (i) *When humans show it:* Unsupervised phonotactic learning (Moreton and Pertsova, 2016; Moreton et al., 2017; Gerken et al., 2019); sometimes in supervised visual concept learning (Kurtz et al., 2013, Exp. 7)
   (ii) *How GMECCS gets it:* Greater neighborhood density (for Type I than IV, and IV than II) $\Rightarrow$ more constraint overlap and hence faster harmony gain (Moreton et al., 2017, §4.1.2).
   (iii) *How Evolutionary Winnow-MaxEnt-Subtree gets it:* Large population, small weight quantum, fitness-insensitive population adjustment $\Rightarrow$ Finds one-feature approximations to Types I and IV, but more-complex constraints needed for Type II take much longer.

b. $I > II > IV$
   (i) *When humans show it:* Supervised visual concept learning, when rule-seeking is encouraged (Shepard et al., 1961; Nosofsky et al., 1994; Smith et al., 2004; Kurtz et al., 2013)
   (ii) *How serial hypothesis-testing models get it:* Hard-wired bias in favor of hypotheses that use fewer features
   (iii) *How Evolutionary Winnow-MaxEnt-Subtree gets it:* Small population, large weight quantum, fitness-sensitive population adjustment $\Rightarrow$ invalid constraints gradually become unavailable, and a valid grammar becomes more and more likely to arise by chance.

(38) How else might evolutionary ideas be applicable in phonological learning? Some possibilities:

a. *Abruptness* is a familiar aspect of first-language acquisition ("across-the-board" changes, e.g., Smith 1973; Macken and Barton 1978; Vihman and Velleman 1989; Barlow and Dinnsen 1998; Levelt and van Oostendorp 2007; Gerlach 2010; Becker and Tessier 2011; Guy 2014), and has been observed in lab-learned phonology (Moreton and Pertsova, 2016). It has often been interpreted as diagnostic of serial hypothesis testing (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. Evolutionary Winnow-MaxEnt can show similar behavior while searching for a constraint. Example (from Moreton 2019):

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.)



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).

b. *Priming* of new constraints by old ones. More-successful macro-constraints spawn more mutants, causing the constraint space around themselves to be searched more intensively. This could steer language change so as to cause a language to idiosyncratically re-use the same features across multiple patterns in its grammar (see Carter 2017 for some relevant findings).

c. *Evolving candidates* from an input to find most-harmonic output — a sort of evolutionary version of Serial Harmonic Grammar (McCarthy, 2010).

d. *Storing instances* to make lexically-specific constraints (Pater, 2009; Moore-Cantwell and Pater, 2016). Those can then evolve towards more generality — a sort of evolutionary version of exemplar theory (Hintzman, 1984, 1986; Bybee, 2001; Pierrehumbert, 2001, 2002; Kirchner et al., 2010; Pierrehumbert, 2016).

(39) Some reasons for skepticism (about this model, not about evolutionary computing in general):

a. Evolutionary Winnow-MaxEnt-Subtree seems to predict that gradual weight update and serial hypothesis testing are mutually exclusive, at least at any particular time. (Potential fix: Test to see if they are.)

b. Many combinations of parameter settings yield poor learning. (Potential fix: Simplify the model to shrink parameter space.)

c. Learning phonotactics from winner-loser pairs, or even n-tuples, is suspect in terms of cognitive realism — generating well-formed nonsense words is not something most people do without practice (e.g., scat singing, glossolalia). Potential fix: pending.

References

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.

Bäck, T. (Ed.) (1996). *Evolutionary algorithms in theory and practice : evolution strategies, evolutionary programming, genetic algorithms.* New York: Oxford University Press.

Barlow, J. A. and D. A. Dinnsen (1998). Asymmetrical cluster development in a disordered system. *Language Acquisition 7*(1), 1–49.

Bateman, N. (2007). *A crosslinguistic investigation of palatalization.* Ph. D. thesis, University of California, San Diego.

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

Boersma, P. (1997). How we learn variation, optionality, and probability. MS, Rutgers Optimality Archive.

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

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

Bybee, J. (2001). *Phonology and language use.* Cambridge, England: Cambridge University Press.

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.

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.

De Jong, K. A. (2006). *Evolutionary computation: a unified approach.* Cambridge, Massachusetts: MIT Press.

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.

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

Endress, A. D. and L. L. Bonatti (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition 105*(2), 247–299.

Endress, A. D. and J. Mehler (2009). Primitive computations in speech processing. *Quarterly Journal of Experimental Psychology 62*(11), 2187–2209.

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

Futrell, R., A. Albright, P. Graf, and T. J. O'Donnell (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics 5*, 73–86.

Gerken, L., C. Quam, and L. Goffman (2019). Adults fail to learn a type of linguistic pattern that is readily learned by infants. *Language Learning and Development*.

Gerlach, S. R. (2010). *The acquisition of consonant feature sequences: harmony, metathesis, and deletion patterns in phonological development.* Ph. D. thesis, University of Minnesota.

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*(3), 227–247.

Goddard, I. (1979). *Delaware verbal morphology: a descriptive and comparative study.* Taylor and Francis.

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

Golston, C. (1996). Direct Optimality Theory: Representation as pure markedness. *Language 72*(4), 713–748.

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.

Guy, G. R. (2014). Linking usage and grammar: generative phonology, exemplar theory, and variable rules. *Lingua 142*, 57–65.

Hebert, R. J. and N. Poppe (1963). *Kirghiz manual*, Volume 33. Bloomington: Indiana University Press.

Hintzman, D. L. (1984). MINERVA 2: a simulation model of human memory. *Behavior Research Methods, Instruments, and Computers 16*(2), 96–101.

Hintzman, D. L. (1986). "schema abstraction" in a multiple-trace memory model. *Psychological Review 93*(4), 411–428.

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.

Kapatsinski, V. (2011). Modularity in the channel: the link between separability of features and learnability of dependencies between them. Proceedings of the XVIIth International Congress of Phonetic Sciences.

Kirchner, R., R. K. Moore, and T.-Y. Chen (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics 38*(4), 540–547.

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.

Kronfeldner, M. E. (2010). Darwinian "blind" hypothesis formation revisited. *Synthese 175*, 193–218.

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.

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.

Levelt, C. and M. van Oostendorp (2007). Feature co-occurrence constraints in L1 acquisition. *Linguistics in the Netherlands 24*(1), 162–172.

Lin, Y. (2009). Tests of analytic bias in native Mandarin speakers and native Southern Min speakers. In Y. Xiao (Ed.), *21st North American Conference on Chinese Linguistics*, Smithfield, Rhode Island, pp. 81–92. Bryant University.

Litman, G. W., J. P. Rast, and S. D. Fugmann (2010). The origins of vertebrate adaptive immunity. *Nature Reviews Immunology 10*, 543–553.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning 2*, 285–318.

Löfstedt, I. (1992). Swedish segment length and Structure Preservation. *Studia Linguistica 46*(2), 93–127.

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

Luquiens, F. B. (1909). *An introduction to Old French phonology and morphology.* New Haven, Connecticut: Yale University Press.

Macken, M. A. and D. Barton (1978, March). The acquisition of the voicing contrast in English: a study of voice-onset time in word-initial stop consonants. Report from the Stanford Child Phonology Project.

Maddox, W. T. and F. G. Ashby (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes 66*, 309–332.

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

McCarthy, J. J. (2010). An introduction to Harmonic Serialism. MS, University of Massachusetts, Amherst.

Mielke, J. (2008). *The emergence of distinctive features.* Oxford, England: Oxford University Press.

Moore-Cantwell, C. and J. Pater (2016). Gradient exceptionality in maximum entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics 15*, 53–66.

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

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. (2019). Constraint breeding during on-line incremental learning. In *Proceedings of the Society for Computation in Linguistics*, Volume 2, pp. Article 9.

Moreton, E. (2020). Evolving constraints and rules in Harmonic Grammar. In *Proceedings of the Society for Computation in Linguistics*, Volume 3, pp. Article 8.

Moreton, E., J. Pater, and K. Pertsova (2017). Phonological concept learning. *Cognitive Science 41*(1), 4–69.

Moreton, E. and K. Pertsova (2014). Pastry phonotactics: is phonological learning special? In H.-L. Huang, E. Poole, and A. Rysling (Eds.), *Proceedings of the 43rd Annual Meeting of the Northeast Linguistic Society, City University of New York*. Amherst, Massachusetts: Graduate Linguistics Students' Association.

Moreton, E. and K. Pertsova (2016). Implicit and explicit processes in phonotactic learning. In TBA (Ed.), *Proceedings of the 40th Boston University Conference on Language Development*, Somerville, Mass., pp. TBA. Cascadilla.

Nakano, K., H. Hiraki, and S. Ikeda (1995). A learning machine that evolves. In *Proceedings of ICEC–95*, pp. 808–813.

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.

Nosofsky, R. M., T. J. Palmeri, and S. C. McKinley (1994). Rule-plus-exception model of classification learning. *Psychological Review 101*(1), 53–79.

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. 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.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*, pp. 137–157. Amsterdam: John Benjamins.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven and N. Warner (Eds.), *Papers in Laboratory Phonology VII*, pp. 101–140.

Pierrehumbert, J. B. (2016). Phonological representation: beyond abstract versus episodic. *Annual Review of Linguistics 2*, 33–52.

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. 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.

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.

Smith, N. V. (1973). *The acquisition of phonology: a case study.* Cambridge, England: Cambridge University Press.

Vihman, M. M. and S. Velleman (1989). Phonological reorganization: a case study. *Language and Speech 32*, 149–170.

Weinert, S. (2009). Implicit and explicit modes of learning: similarities and differences from a developmental perspective. *Linguistics 47*(2), 241–271.