2012

Structurally Biased Phonology: Complexity in Learning and Topology

Joe Pater
Elliott Moreton, *University of North Carolina at Chapel Hill*

JOE PATER AND ELLIOTT MORETON

STRUCTURALLY BIASED PHONOLOGY: COMPLEXITY IN LEARNING AND TYPOLOGY*

ABSTRACT**:** This paper presents structurally biased phonology, a program of research that aims to formalize and better understand the role of structural complexity in phonological learning and typology. The paper situates the program with respect to other research in generative phonology, and provides a framework for it, termed Incremental MaxEnt with a Conjunctive Constraint Schema (IME/CCS). This framework extends previous proposals in generative phonology, cognitive psychology and machine learning to the study of structural complexity. IME/CCS is successfully applied to model some illustrative cases in which structurally simpler patterns have been found easier for humans to learn. It is also shown to make predictions about skews toward simplicity in typology, in conjunction with a model of iterated learning.

KEYWORDS: *Simplicity, complexity, learning bias, inductive bias, feature economy, regularization, probabilistic grammar, optimality theory, harmonic grammar, maximum entropy, gradual learning algorithm, structurally biased phonology, iterated learning, agent-based modeling, artificial language learning*

---

* Portions of this research were presented at the University of North Carolina Chapel Hill, April 2008, the Boston University Conference on Language Development and NECPhon at Yale University, both November 2008, the KNAW Academy Colloquium on Language Acquisition and Optimality Theory, Amsterdam, July 2009, the Cornell Workshop on Grammar Induction, May 2010, the meeting of the Réseau Français de Phonologie, Orléans, July 2010, the MOT conference on Phonology in the 21st Century: In Honour of Glyne Piggott, Montreal, May 2011, the LSA Workshop on Testing Models of Phonetics and Phonology, Boulder, July 2011, and at the Chicago Linguistics Society, April 2012. We especially thank Michael Becker and Katya Pertsova for their collaboration on the BUCLD/NECPhon and CLS presentations respectively, as well as the organizers of these events. We also received helpful comments at these venues and elsewhere from Adam Albright, Seth Cable, Andrew Cohen, Gary Dell, Brian Dillon, Elan Dresher, Minta Elsman, Robert Frank, Matt Goldrick, Gaja Jaroz, Mark Johnson, John Kingston, John McCarthy, Claire Moore-Cantwell, Alex Nazarov, Presley Pizzo, Lisa Sanders, Amanda Seidl, Brian Smith, Jen Smith, Paul Smolensky, Robert Staubs, and Colin Wilson. This research was supported by NSF Grant 0813829 to the University of Massachusetts Amherst.

The EFL Journal 3:2 June 2012.

0. INTRODUCTION: THE PROGRAM

Structurally biased phonology is a program of research that aims to understand the role of structural complexity in phonological learning and typology. Lab experiments have repeatedly found that the difficulty of learning a pattern increases along with the number of phonological features on which it crucially depends (reviewed in Moreton & Pater to appear). Among natural languages, featurally-complex inventories and phonological patterns are also under-represented relative to featurally simple ones (Clements 2003; Mielke 2004, Ch. 6; Moreton 2008; Mackie & Mielke 2011). As we explain in what follows, current phonological theory does not provide an account of either of these effects of structural complexity. Although a form of structurally biased phonology was originally proposed in Bach and Harms (1972), it seems not to have been subsequently much pursued. We seek to better understand: (1) what effects structure has on learning and typology, (2) whether and how the learning effects are causally connected with the typological ones, (3) what properties a learner must have to account for them, and (4) whether and how these effects are connected to other domains in and out of linguistics.

The main goal of this paper is to present a theoretical framework for this research, including some initial illustrative results. In the sections that follow this one, we present Incremental MaxEnt with a Conjunctive Constraint Schema (IME/CCS; suggested pronunciation ['aimɛks]) using some relatively simple examples of the modeling of learning and typology (section 1), before turning to some somewhat more complex modeling (section 2). These simulations illustrate the general predictions that our model makes for both learning and typology. The framework is largely based on previous proposals in linguistics, and its component parts have also seen broad application in cognitive psychology and machine learning. The novelty here is in its use to derive structural biases in phonology (though see Martin 2011 for related work).

In this first section 0, we place the general program of research within the context of other programs of research currently being pursued in phonology. Our aim in doing so is not only to contextualize and to hopefully stimulate interest in the particular framework we have been developing, but also to inspire the construction of competing frameworks for research in structurally biased phonology.

0.1 Simple examples of structural complexity

To set up the discussion, we start with a pair of examples of the role of structural complexity in typology and learning. The following typological example is due to John Kingston (p.c.). The tables in (1) provide examples of obstruent stop inventories that differ in how they use the [+/–voice] feature across places of articulation.

(1) *Patterns of [+/–voice] contrast across place*

A

	[Lab]	[Cor]	[Dor]
[–voice]	p	t	k
[+voice]			

B

	[Lab]	[Cor]	[Dor]
[–voice]	p	t	k
[+voice]	b	d	g

C

	[Lab]	[Cor]	[Dor]
[–voice]	p	t	k
[+voice]	b	d	

Inventory A has a simple description in phonetic terms: voiceless labial, coronal, and dorsal stops. Inventory B is only slightly more complicated: voiced and voiceless labial, coronal, and dorsal stops. But Inventory C's description is more complex, since it requires stipulations about which feature values can co-occur: voiced and voiceless labial and coronal stops, and voiceless dorsal stops. The same is true of any other inventory that has [p, t, k] and some but not all of [b, d, g].

Natural-language inventories tend to be like A or B, not C. Table 1 is based on those languages in the genetically and areally balanced UPSID-92 database (Maddieson & Precoda 1992) whose inventories have all of [p, t, k]. Of these, 244 have both [b] and [g], 153 have neither [b] nor [g], 43 have only [b], and 11 have only [g]. The Observed/ Expected ratios compare the actual counts to those that we would expect if [b] and [g] occurred independently of each other. Clearly, they are not independent: Inventories which have [b] tend to also have [g], and vice versa (chi-squared = 257, df = 1, $p < 0.001$).

(2) *Distribution of voicing contrasts by place in UPSID-92*

	[b]	no [b]
[g]	244 (O/E = 1.52)	11 (O/E = 0.12)
no [g]	43 (O/E = 0.34)	153 (O/E = 2.13)

This cross-linguistic tendency for inventories to avoid "bumps" and "holes" is documented in Clements (2003) and subsequent work (Clements 2009, Mackie & Mielke 2011).

Saffran and Thiessen (2003) provide experimental results on learning by 9-month-olds that show that the class of segments defined by just [+/–voice] is easier to learn than a class that requires both [+/–voice] and place of articulation to define. The experiment involved both learning of a phonotactic pattern and word segmentation. In the first training phase the infants were exposed to isolated words of shape $C_1V\ C_2.C_1V\ C_2$, where C_1 and C_2 were each limited to a set of three consonants. In the second training phase, they were exposed to 4 new words in a continuous stream, with only two fitting the pattern from the first training phase. In subsequent testing, listening time was measured (using visual fixation) for each of the 4 words from the second training phase, presented in isolation. When the pattern was featurally simple, with voiceless [p, t, k] in one position, and voiced [b, d, g] in the other, the infants displayed a novelty preference for non-conforming items. When the pattern was featurally complex, with [p, d, k] in one position, and [b, t, g] in the other, there was no significant difference in listening time between conforming and non-conforming items.

0.2 Structure and substance in generative phonology

These and similar data from learning and typology have yet to have been given an account. Chomsky and Halle's (1968:334) evaluation procedure prefers grammars that use fewer features, but it was originally proposed only as a means of choosing between analyses of a single set of data, and does not make one pattern easier to learn than another, or predict that one pattern should be more common typologically (though see Bach and Harms 1972 for a possible extension). It is possible that a

learning algorithm incorporating the evaluation procedure or some other Minimum Description Length principle could be used to account for simplicity biases in learning and typology, but this remains to be shown. Complex patterns can be learned, just with more difficulty, and are only typologically under-represented, not absent. These sorts of probabilistic tendencies fall out of the scope of most current phonological theories.

Generative phonology, as it has developed from Chomsky and Halle (1968) onwards, aims to provide a formal characterization of the space of possible phonological systems. This research program has yielded a wealth of information about the structure of phonological systems, and a wide range of useful formalizations of this structure. The dominant framework for this research in the last twenty years has been Optimality Theory (OT; Prince & Smolensky 1993/2004), which makes use of violable constraints to analyze individual languages and to generate typological predictions. Even though OT incorporates violable constraints, its typological predictions are still *all-or-nothing* – either a language is generated by a set of constraints, or it is not (though see Coetzee 2002 and Bane & Riggle 2008 for possible extensions). Other theories of phonology are sometimes claimed to make more gradient predictions. For example, in a review of pre-OT research on feature geometry, McCarthy (1988:84) states that "[t]he goal of phonology is the construction of a theory in which cross-linguistically common and well-established processes emerge from very simple combinations of the descriptive parameters of the model". Similar statements about a relationship between simple feature geometric formalisms and typologically common patterns are also found in Clements (1985) and Sagey (1990). However, in all cases, the causal mechanism linking the formalism and typology is left completely unspecified. We discuss other versions of this issue in the conclusions of 1.2 and 2.2, after we have shown how learning can provide the link from formal grammatical structure to gradient typology.

The fact that phonological frameworks only make categorical predictions about typology is just one reason why they require amendments to serve as frameworks for structurally biased phonology. Another issue is that they have typically conflated what are called structural and substantive factors, and these may well be the products of separate systems. Before

turning to the problem, it's worth recognizing that there are good arguments for incorporating substance into a theory of phonological typology. From the outset of generative phonology, it was realized that a purely formal system would not distinguish between common phonological patterns and unattested ones (Chomsky & Halle 1968, Ch. 9). It is not merely the structural complexity of a pattern that determines whether it is typologically attested. It is straightforward to change a typologically well-motivated phonological rule or constraint into one that generates undesired results by substituting one feature or variable for another, which keeps complexity constant. For example, a coda devoicing rule becomes a voicing rule by substituting a plus '+' for the minus in [–voice], and a constraint banning voiced codas can as just easily become one that bans voiceless ones. A theory that generates coda voicing is generally held to be undesirable insofar as it fails to fit the attested typology, which lacks this process (see Blevins 2004 for challenges to both the theoretical assumption and the empirical facts; see Kiparsky 2006 for a response).

It is uncontroversial that a full account of phonological typology needs to take into consideration the 'substance' of rules and/or constraints, which usually means their connection to articulation, perception, and perhaps other aspects of language use. What is controversial is how and whether this substance should be incorporated into the phonology itself, that is, into the formal system that is taken to characterize speakers' knowledge of the sound system of a language. Bach and Harms (1972) challenge the notion in Chomsky and Halle (1968, Ch. 9) that the purely formal evaluation procedure should incorporate substantive markedness principles. Anderson (1981) argues against attempts, such as that of Stampe (1979), to define a privileged set of phonetically 'natural' phonological rules (see also Anderson 1985). Ohala (1990) criticizes the idea that formally simple rules should yield common patterns, arguing that the explanations are to be found in the phonetics itself, and in patterns of language change. Finally, Blevins (2004) and Hale and Reiss (2000, 2008) draw on this line of earlier critique in questioning the use of substantive factors in determining the contents of OT's universal constraint set (see especially Hayes, Kirchner & Steriade 2004 on explicitly phonetically driven versions of OT, and Archangeli & Pulleyblank 1994 for a pre-OT precedent).

There are two main worries about substantively grounded theories of phonology that permeate these critiques, worries that seem to us legitimate. The first is that there are well-documented instances of productive phonological patterns that do not have a synchronic phonetic basis (see e.g. Icelandic velar fronting in Anderson 1981, NW Karaim consonant harmony in Hannson 2007, and Sardinian [l] ~ [ɾ] alternations in Scheer forthcoming).[1] The second is that it seems in principle possible to derive the fact that many phonological patterns are phonetically grounded from their diachronic emergence from the phonetics itself – see Yu (to appear) for a recent collection of papers on 'phonologization'.

The first of these concerns might be met by developing a substantively *biased* framework for phonology (Wilson 2006). Instead of adopting substantively grounded phonology's hard requirement that phonological rules or constraints be phonetically motivated, which makes ungrounded phonology unlearnable, one might instead build a theory in which grounding facilitates learning (this is also the intent of Chomsky and Halle 1968, Ch. 9 and Stampe 1979). Wilson (2006) motivates such a theory on the basis of artificial language learning experiments that are claimed to provide evidence of greater generalization of phonetically grounded patterns. Our own reading of the results of Wilson's experiments, as well as other related ones, is that the evidence for substantive bias is weak at best, especially in comparison with the strong evidence from this experimental paradigm for structural biases like that seen in the Saffran and Thiessan experiment reviewed above (see Moreton & Pater 2012 for extended discussion). Based on current evidence, it remains plausible that many of the substantive skews in phonological typology, especially the ones most clearly related to phonetics, do arise in phonologization, rather than being encoded in a phonological inductive bias. We follow Wilson in seeking a framework that incorporates inductive bias in learning, but the biases that we seek are structural, rather than substantive (though see further section 1.1).

[1] Kiparsky (2006) and Kingston and de Lacy (to appear) provide examples of patterns that they claim are ruled out by Universal Grammar (UG), and which cannot be produced by patterns of historical change. It remains to be seen whether a fully explicit theory of UG can deliver these results, yet still allow for the acquisition of patterns like the ones we have listed in the text (see also Hayes, Zuraw, Siptar & Londe 2009 for further challenges).

Given the concerns that have been expressed about substantively grounded phonology, why has it been the focus of such sustained research activity? There are likely as many answers as there are phonologists, but we can see a few reasons why alternatives have not yet been sufficiently explored. The main one is that as we have mentioned above, some type of substantive grounding has been crucial to the success of all theories of phonology as theories of typology. To take the case of OT, the typological predictions of the theory rely on the presence *vs.* absence of constraints from its universal constraint set, and often constraints that are formally identical lie on opposite sides of this divide. For example, Prince and Smolensky's (1993/2004) syllable structure typology relies on the presence of constraints demanding onsets and banning codas, and the absence of constraints that ban onsets and demand codas. There is nothing that formally distinguishes ONSET and NOCODA from NOONSET and CODA: there is no general formal property that picks out the first two as special. They are only distinguished by their substance, that is, by the particular configuration of structural elements involved. This is not a peculiarity of OT – the parametric theories from which it descended have similar stipulations about the contents of their sets of parameters.

There is a body of current research pursuing alternatives to substantively grounded phonology, but it has yet to address the role of structural complexity in learning in typology. One approach pursues purely formal theories of phonology, but leaves the learning component unspecified (e.g. Hale & Reiss 2008, Samuels 2011). The result is that these theories make no predictions about the effects of structural complexity in learning, nor do they have anything concrete to say about its effects on typology.[2] Another approach is to study how phonetics and diachrony can explain skews in phonological typology. A prominent example of this research is Blevins (2004), which does push that program of research forward, but contains no explicit proposal about the structure of the phonological component, and says nothing concrete about the role of phonological complexity in learning and typology.

[2] Hale and Reiss (2008) and Samuels (2011) eschew typological restrictiveness as an evaluation metric on theories, but it is not clear what replaces it.

In sum, current phonological theory seems to be at an impasse. Typological evidence points to the role of substantive factors in shaping phonology, yet individual phonological systems are apparently content to maintain ungrounded patterns (and even innovate them, according to Bach & Harms 1972). Based on the experimental evidence surveyed in Moreton and Pater (to appear), it also seems that learners have no problem in acquiring phonetically ungrounded systems. The best way forward, it seems to us, is to take on the difficult task of constructing theories of phonological typology in which the phonological component itself is not directly constrained by the phonetics. This task is difficult because it involves constructing interacting theories of phonetics, phonology, learning, and change, and measuring the outcome of this interaction against typology. The empirical work involved in this sort of research program also raises its own set of challenges: how do we go about teasing apart the effects of inductive bias in phonological learning from the effects of phonetic transmission? One answer to this question is provided in Moreton (2008), where the study of phonetics and artificial phonology learning is used to argue that a set of typological skews are the joint product of a structural inductive bias (*aka* analytic bias) and the effects of phonetic transmission (*aka* channel bias – see Yu 2012 for discussion of whether this was measured properly).

0.3 Structurally biased phonology in generative phonology

In this paper, we will not take on the full task of showing how a structurally biased phonology can jointly produce typological distributions with theories of phonetics and learning. Instead, we take on the smaller, yet still challenging, job of developing a framework that can make predictions about the role of structural complexity in learning and typology. We see this as only one piece of a much larger and quite complicated system, whose dynamics we will only be able to understand by building explicit models of each part. Here we sketch how the model we develop relates to previous streams of research in generative phonology.

Our research in structurally biased phonology is firmly within the generative tradition in its concern with explicit formal description and typological explanation, but departs from much research within it in a number of ways. The first is that we adopt probabilistic grammar

formalisms, which have only recently become popular within 'mainstream' phonology (see Coetzee & Pater 2011 for discussion). Probabilistic grammars are useful, maybe even indispensible, in modeling the course of learning and of language change (see e.g. Jarosz 2010 for an overview of models of the development of phonological production that use probabilistic grammars, and Zuraw 2003 on probabilistic models and diachrony).

The biggest difference between structurally biased phonology and most generative research is in that it seeks to formalize soft, as well as hard restrictions on learning and typology. The adoption of probabilistic grammar models does not in itself automatically deliver accounts of probabilistic typological tendencies such as the tendency towards feature economy. To create such accounts, we draw on a line of research that largely sets itself outside of the generative tradition, which involves the creation of computational models of agent-based or iterated learning to model change and probabilistic typological generalizations (see Wedel 2011 for an overview; see Hare & Elman 1995 for the pioneering application to structural simplicity in phonology). Much of the work on iterated learning, including Wedel's own, uses analogical mechanisms to capture agent-internal generalizations that would be attributed to grammars in generative phonology. We are impressed by the general success of grammar-based research on language, and hence adopt grammatical rather than analogical models, but their relative merits within an iterated learning framework deserve further scrutiny. In the work presented below, we proceed by identifying learning biases that emerge from a model of grammar and its learning, and by then generating typological probabilities through repeated runs of iterated learning using this model.

The final distinctive feature of structurally biased phonology with respect to at least the bulk of generative research is its use of laboratory methods to study learning biases. While the rise of Laboratory Phonology (starting with Kingston & Beckman 1990, see Pierrehumbert *et al*. 2000) has made experimental methods a core component of phonological practice, it is still the case that most proposals about the contents of a phonological UG come from the traditional method of constructing a theory that comes

as close as possible to generating all and only the patterns attested amongst the world's languages. The laboratory study of learning is a powerful, if undoubtedly imperfect, way of teasing apart inductive biases such as UG from the typological effects of phonetic transmission. There are two other methods that we are aware of, but that we will not make use of in this paper. One is to conduct experiments (e.g. "wug"-tests, nonce word judgments) with speakers of naturally learned languages to find out how their internalized knowledge corresponds to patterns in the hypothesized learning data (for data suggestive of simplicity biases, see Hayes *et al.* 2009 and Becker *et al.* 2011). Another is to construct explicit models of phonetic transmission and learning, in order to see what is left over for phonology to explain (see Wedel 2011 and Yu to appear for references and a recent collections of papers). An important precedent in this latter area to our own work is that of Boersma and Hamann (2008) and Boersma (2011), because the model of grammar and learning developed there is so close to ours, to which we will now turn. The main difference is a matter of focus: Boersma and colleagues have not studied the structural biases that are our concern.

1. A FRAMEWORK: INCREMENTAL MAXENT WITH A CONJUNCTIVE CONSTRAINT SCHEMA

The *desiderata* for a framework for structurally biased phonology are quite different from those for many other approaches to phonology. Most fundamentally, within this framework we will not expect a theory of some domain, like vowel harmony, to generate only the patterns observed cross-linguistically, since we expect typological generalizations to emerge from the interaction of theories of phonology, phonetics, learning and change. Instead, we require models that will deliver just the preferences for structural simplicity observed in learning, and that will predict skews toward structural simplicity in typology. Because the aims of structurally biased phonology are different from other programs of research in phonology, it would be surprising if a successful framework for it directly resembled a framework designed for other purposes. Because the empirical domain of structurally biased phonology overlaps considerably

with that of other approaches, it would also be surprising if there were nothing to be gained here from importing formalisms from preceding theoretical frameworks. Our own proposals build on research in OT by using violable constraints, though we depart in several ways from the framework presented in Prince and Smolensky (1993/2004). The primary usefulness of violable constraint grammars in the present context is that they have associated well-developed learning algorithms. It turns out that a bias for structural simplicity, the basic desired property of theories of structurally biased phonology, falls out directly from relatively minimal assumptions about the nature of the grammar and learning algorithm.

As mentioned in the preceding section, our framework uses probabilistic grammar models, as opposed to the categorical one proposed in Prince and Smolensky (1993/2004). Here we adopt maximum entropy or MaxEnt grammars (Goldwater & Johnson 2003, Hayes & Wilson 2008). Unlike the original proposals just cited, we combine these models with an on-line, gradual learning algorithm, which is suitable for modeling the course of learning. The combination of a MaxEnt grammar with an incremental learner was first explored in generative phonology by Jäger (2007), who calls the learning algorithm a sampling version of Stochastic Gradient Ascent. Here, we follow Johnson (2007) and Pater (2008) in referring to this learning algorithm as the Perceptron update rule (after Rosenblatt 1958), leaving Stochastic Gradient Ascent for the non-sampling version (see Johnson 2007 for the formalizations). This is the same learning algorithm that Boersma and Pater (to appear) call HG-GLA, and it's also referred to as the Delta Rule in connectionist research. Because we use incrementally learned MaxEnt grammars, we refer to our framework as Incremental MaxEnt, though it's likely that many of our results could also be obtained with other gradually learned probabilistic grammar models, like Boersma's (1997) incrementally learned stochastic version of OT.

A MaxEnt grammar uses weighted constraints to define a probability distribution over a set of representations. We use two kinds of these grammars in our simulations. In the type proposed by Goldwater and Johnson (2003), the probability distribution is defined over the candidate set of outputs for a given input (e.g. over candidate phonological surface

representations for a given underlying representation). We also follow Hayes and Wilson (2008) in using a MaxEnt grammar that defines a probability distribution over a space of possible words, that is, as a probabilistic model of phonotactics. In either case, the probability of a representation is proportional to the exponential of its weighted sum of constraint scores. Constraint scores can be violations or rewards – we use rewards for convenience in our simulations. The weighted sum of scores is the quantity termed Harmony in Harmonic Grammar (Smolensky & Legendre 2006, Pater 2009), so we can say that probability is proportional to $\exp(H)$.

The other fundamental assumption of our framework is that the constraint set has a particular structure, which follows from the application of a conjunctive constraint schema to a given set of features; hence the framework as a whole is dubbed Incremental MaxEnt with a Conjunctive Constraint Schema (IME/CCS). We will introduce this assumption in the context of the simulation that follows.

1.1 Learning with a MaxEnt phonotactic grammar

To show how a simplicity bias can be straightforwardly obtained with these assumptions about grammar and learning, we'll use a simplified version of the learning scenario in the Saffran and Thiessen (2003) experiment described above. We assume as a space of possible representations the 6 consonants in Table 2: voiced and voiceless labials [b] and [p], coronals [d] and [t], and dorsals [ɡ] and [k]. The sum of the probabilities within this representational universe adds up to 1. We'll see that a learner that starts with probability equally distributed amongst the consonants will shift the probability to observed [p, t, k] more quickly than to observed [p, d, k], that is, it will learn the [p, t, k] distribution more quickly than [p, d, k].

The table in (3) uses just three constraints to show how a MaxEnt phonotactic grammar works. The first two constraints [+Vce] and [–Vce] reward consonants that are voiced and voiceless respectively, assigning a score of +1 in each case. The third constraint [+Vce]$\wedge$[+Cor] rewards a conjunction of features and thus assigns +1 to the voiced coronal [d]. The weights of the constraints are shown in the first row

underneath their names, and each consonant's weighted sum of scores, or Harmony, is shown in the column labeled *H*. For example, the *H* score for the voiced coronal is $(1 \times -4) + (1 \times 8) = 4$. The probabilities result from dividing exp(*H*) for each consonant by the sum of all 6 of these numbers. With these constraint weights, the probability is roughly equally divided between [d], [p], [t] and [k], with a vanishing amount of probability reserved for [b] and [g]. Negatively weighted [+Vce] and positively weighted [–Vce] both shift probability onto the voiceless consonants, while positively weighted [+Vce]∧[+Cor] allows [d] to retain a non-negligible proportion of the probability mass.

(3) *A MaxEnt phonotactic grammar*

	[+Vce] –4	[–Vce] 4	[+Vce]∧[+Cor] 8	*H*	*P*
[b]	1			–4	< 0.001
[d]	1		1	4	0.25
[g]	1			–4	< 0.001
[p]		1		4	0.25
[t]		1		4	0.25
[k]		1		4	0.25

We apply the Perceptron update rule to MaxEnt phonotactics as follows. Each learning step begins by sampling a single learning datum from the target distribution. An example of a target of learning within our 6 consonant universe is a uniform probability distribution of 0.33 for each of the voiced consonants, with voiceless consonants having probability 0. Sampling from this distribution might yield [d]. Next, a representation is sampled from the distribution defined by the current grammar. Assuming the grammar in Table 1, there is a 25% chance that we would sample [p]. We then take the difference in the scores of the two representations, which yields the vector (1, –1, 1) for the constraints in Table 1. This vector is scaled by the learning rate, and the result is added to the current weights to get the updated weight values. This would here increase the weights of [+Vce] and [+Vce]∧[+Cor], and decrease that of [–Vce], thus shifting probability from [p] to [d].

For our simplified simulation of the Saffran and Thiessen results, we model the learning of two target distributions over these 6 consonants.[3] In the *ptk* language, the voiceless consonants each have 1/3 of the probability, and the voiced ones have none. In the *pdk* language, the voiceless labial and dorsal each have 1/3 of the probability, as does the voiced coronal, while the rest of the consonants have probability 0. For each of the languages, the constraint set consists of the constraints that reward each observed single feature, as well as each two-feature conjunction. We thus have the constraint sets in (4). We have not included in (4) the single feature constraints for the 3 place features, which were included in the simulations, but played no role in the results.

(4) *Constraint sets for learning simulation*

a. *ptk* language: [–Vce], [+Lab]∧[–Vce], [+Cor]∧[–Vce], [+Dor]∧[–Vce]

b. *pdk* language:[–Vce], [+Vce], [+Lab]∧[–Vce], [+Cor]∧[+Vce], [+Dor]∧[–Vce]

The constraint weights started at zero, and the learning rate was set at 0.01. The graph in Figure 1 shows the probability assigned to the observed forms over the course of 2000 learning trials, at 50 trial intervals. These probabilities are the averages over 10 runs. At every point in learning, the *ptk* learner assigns higher probability to the observed forms in its language than the *pdk* learner does.

[3] All simulations were run as scripts in R (R development core team 2010). The basic script, which can be straightforwardly modified to replicate the simulations, can be found at http://blogs.umass.edu/hgr/perceptron/.

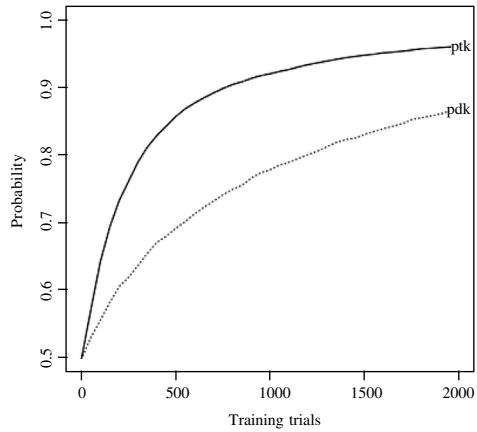

Figure 1: Probability assigned to observed forms

The following three figures show why *ptk* is easier to learn than *pdk* under these assumptions. First, for every mismatch between the target datum and the one generated by the current grammar, the *ptk* learner adds positive weight to [–Vce]. This constraint thus rises quickly, as shown in this graph of constraint weights over the first 500 trials. The [–Vce] constraint is labeled *vcl* in the graph, while the two-feature constraints are labeled with the consonants they reward: [p], [t], and [k]. The place constraints remain around zero, since the target probability distribution is uniform with respect to place.

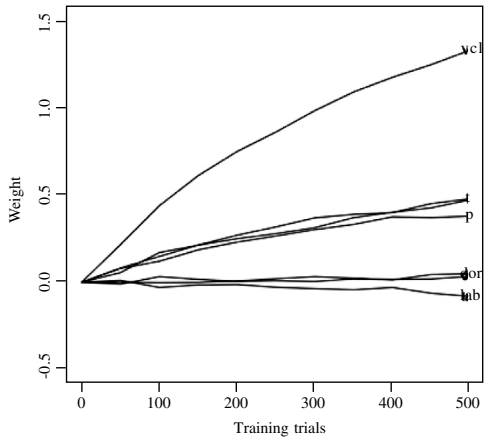

Figure 2: Constraint weights, first 500 trials, *ptk* condition

In contrast, there is no single constraint that picks out the observed forms for the *pdk* learner. Because voiceless consonants predominate, [–Vce] does get a positive weight, and [+Vce] (*vcd*) a negative weight. Their joint effects must be overcome by a high weighted [+Cor]∧[–Vce] (*d*) constraint (a similar configuration is shown in table (3)).

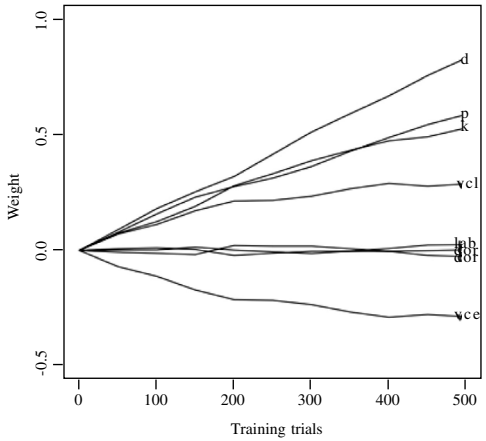

Figure 3: Constraint weights, first 500 trials, *pdk* condition

The weights of the voicing constraints lead to some over-generalization in the early stages of learning. As shown in Figure 4, the learner assigns higher probability to [t] than to [d] in trials 50 through 200, even though it never sees [t].

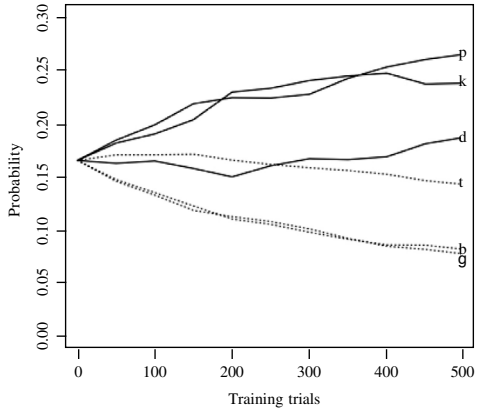

Figure 4: Probabilities of consonants, *pdk* condition

This over-generalization is reminiscent of the behavior of connectionist models (see e.g. Rumelhart & McClelland 1986), and has a similar source, since the model we have presented here is essentially a notational variant of a single layer feedforward network trained with the same update rule (see further the discussion of Gluck & Bower 1988ab in section 2.1).

This grammar and learning model displays a simplicity bias in that the pattern that can be described with a single voicing feature is learned faster than the one that requires both place and voicing. This bias was not written into the model in terms of a prior over constraint weights, as in Wilson's (2006) substantively biased phonology (see Johnson 2007 on how to formalize priors as decay terms in on-line models; see also Jesney and Tessier's 2011 use of learning rate and initial weights). The key assumptions are (1) that the learner's internal vocabulary for representing generalizations consists of constraints that are stated in terms of conjunctions of one or more phonological feature values, and (2) that the contribution of each constraint to the learner's grammar is only adjusted on learning trials for which that constraint is relevant. If a pattern can be stated with reference to a few general conjunctive classes, it will be learned fast, since the constraints that support it will be both relevant and right on many training trials, and will therefore advance quickly. If a pattern can only be stated with reference to many particular classes, it will be learned slowly, since the correct constraints (a) being relevant on just a few training trials, will advance slowly, and (b) can't force a correct decision until they have become strong enough to outvote the incorrect general constraints which, being relevant on many trials, will gain strength quickly but be wrong about the details of the pattern (like positively weighted [–Vce] and negatively weighted [+Vce] in the *pdk* learner).

Thus the only necessary assumption, besides those included in the basic structure of the framework, regards the structure of the constraint set. A complete account requires guaranteeing that the constraint set will contain these general constraints, presumably by incorporating their construction into a constraint induction procedure (see Hayes & Wilson 2008 on constraint induction for non-incremental MaxEnt; see Moreton 2010 on on-line constraint induction). The simplicity bias requires the general single feature constraint [–Vce]; an alternative model that

contains only the two-feature constraints learns both languages at the same rate. The assumption that constraint sets include such general constraints is uncontroversial, but not without content: it implies that learning involves abstraction from the featural make-up of individual forms. For example, there is no observed form that is [–Vce], without also being specified for some place feature. Note too that we are making substantive assumptions about the feature set: we included a feature that groups [p, t, k], and no feature that groups [p, d, k]. In just this way, there is in fact a sort of substantive bias in this theory (see further Moreton and Pater to appear: section 4.4).

Compared to the assumptions that are typically made about constraint sets in OT, our assumptions are quite weak. It remains to be seen to what extent stronger assumptions about the constraint set need to be made to account for phonological learning and typology in this framework, especially once models of phonetics and its learning are incorporated.[4]

1.2 Typology with MaxEnt with candidate sets

To provide a simple illustration how this grammar and learning model can leave an imprint on typology, we turn to a different empirical domain, since feature economy requires a discussion of contrast, which we would like to put off until the next section. The basic specific-to-general relationship between the constraints in the last simulation obtains between many other sets of constraints, and can lead to similar biases in learning. For example, morpheme-specific phonology can be formalized in terms of lexically specific versions of general constraints (Pater 2000, 2010). The constraint set that we posit for this simulation is as in (5):

[4] See Moreton (2012) for evidence that phonological learning requires variables on constraints to account for the privileged status of featural identity (see relatedly Berent *et al.* 2012), and see Staubs (in prep.) on requirements on stress constraints for a typological model of the correlation between main stress position and directionality of footing.

(5) *Constraint sets for typological simulation*

a. Stress-R/L
Assign a reward to if stress is on the right-/leftmost syllable
b. Word-X-Stress-R/L
Assign a reward if stress on Word X is on the right-/leftmost syllable

As in the last simulation, there is a general constraint that applies across contexts, as well as constraints that apply in only specific contexts. Here the contexts are words, rather than places of articulation. We use stress for this example to emphasize the wide scope of our proposals, which are in no way limited to segmental phonology (or even to phonology). The features across which we are applying the conjunctive constraint schema are stress position and Word, with rightmost *vs*. leftmost stress being a stand-in for a more developed representational theory of stress.

The typological skew that we will model is one that most linguists would likely agree exists, and whose existence seems to be a fundamental assumption of linguistic theory, but which remains to be quantified. The skew is towards regularity. That is, we will show that this even though this model of learning and grammar can produce lexical stress, in which stress position is an arbitrary feature of individual words, it predicts a typological skew towards regular stress, where stress position is uniform across words. When one of the general constraints in (5a) has a particularly high weight, stress will be placed uniformly on the right or leftmost syllable. Lexically arbitrary stress arises when stress placement is controlled by the word-specific constraints of the form in (5b). Partially regular systems result from weightings that allow the joint effects of the word-specific and general constraints to be seen.

The typological skew towards regularity is produced by a tendency for grammars to have a high weighted general constraint. Regularization of exceptional stress as seen in in acquisition (e.g. Hochberg 1988), and in diachrony (e.g. Phillips 1984, Sonderegger & Niyogi in press) can be modeled using this sort of constraint set in the present framework; see Pater (to appear) for a demonstration. Here we generate typological predictions by examining the outcome of the interaction of learner-teacher pairs. In each run, two agents repeatedly produce learning data for one

another, with one of the two randomly selected as the teacher in each trial. At the end of a given number of trials, we have a language. A distribution over these languages can be interpreted as the typological prediction of the learning and grammar model. This is not meant to be a realistic model of language use and change; rather, it is a fairly abstract means of generating typological predictions from a theory of both learning and grammar. It abstracts from the possible effects of phonetic transmission, since production and perception are fully accurate, yet it is a step towards greater realism than standard generative models, which consider only the typological effects of a grammar, and not the learning algorithm. This method of modeling typology might be referred to as an Interactive Grammar-Learning model: 'interactive' because it generates predictions from the results of interaction between agents in the simulations, and also because the grammar and the learning algorithm are interacting. This model is in principle independent of our proposals about learning and grammar; one could also implement it with other grammar and learning models.

As in the last simulation, we use a MaxEnt grammar model with the Perceptron update rule, but instead of Hayes and Wilson's (2008) MaxEnt phonotactic model, we use Goldwater and Johnson's (2003) OT-like model. That is, we now have probability distributions over candidate sets for individual words, rather than a single distribution over the space of possible words. There are 4 words, and each has two candidates: one with final/rightmost stress, and one with initial/leftmost stress. There are thus 10 constraints – the two general constraints preferring left- and rightmost stress respectively, as in (5a), and the 4 lexically specific versions of each of these two, as in (5b). The constraints started with zero weight, which produces initial equal probability for the candidates. In each trial, the randomly selected teacher's grammar is used to pick one of the candidate stress positions for a word randomly selected from a uniform distribution. This is then used as the learning datum for the other agent, whose grammar generates the stress position for the same word. These two input-output pairs are used for the Perceptron update described in the last section. Each run consisted of 10,000 learning trials with the learning rate set at 0.1, and there were 50 runs, which produced 50 languages. Any negative weights produced by the update were changed to zero.

The measure of interest is the extent to which stress is in the same position across words. For each word at the end of each run, the candidate with higher probability was taken as the choice of stress position for that word (the two agents always agreed). A baseline probability for uniform stress can be derived as follows. There are 4 words each with a binary choice of stress position, so there are $2^4 = 16$ possible stress position combinations across them. Of these, only $2/16 = 0.125$ have stress in same position – one with all words having leftmost stress, and the having all rightmost. In the outcome of the simulation, $40/50 = 0.80$ of the runs produced uniform stress, a distribution that diverges considerably from chance ($p < 0.001$ by a two-sided exact binomial test).[5]

To provide a more fine-grained view of a subset of the results, we show in (6) the outcome for the first 12 of 50 runs, each time averaged over the two learners. For each row, the column headed 'Word' indicates the word at issue, and 'S' the location of stress, either left- or rightmost. After 10,000 trials, the learners have moved quite far from the uniform 0.50 probability of the initial state. We clearly see the tendency toward regularity: in every run except R9 and R10, the higher probability candidate in every candidate set is uniformly either left stressed (L) or right (R). Note too that the probabilities are tending towards 1 and 0. It may well be of some independent significance that a tendency toward categorical outcomes is emerging from the interaction of learners operating with probabilistic grammar models. Here this tendency is the result of a much-noted outcome of agent-based modeling: the agents are coming to agree on a 'form' for each 'meaning' (see e.g. Liberman 2002 in linguistics, and Schultz *et al*. 2010 in robotics).

[5] Robert Staubs notes that repeatedly assigning random weights to the constraints from a uniform distribution bounded by 0 and 1, without running the learner, produces 40% uniform stress: higher than chance, but lower than the learner and grammar together. Bootstrapping evaluation yields a significant difference ($p < 0.001$) between grammar *vs*. grammar + learner.

(6) *Average probabilities assigned to each stress location*

Word	S	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
1	L	0.04	0.98	0.98	0.00	0.98	0.37	0.02	0.01	0.09	0.85	0.01	0.98
	R	0.96	0.02	0.02	1.00	0.02	0.63	0.98	0.99	0.91	0.15	0.99	0.02
2	L	0.03	0.99	0.66	0.03	0.97	0.05	0.17	0.02	0.25	0.23	0.01	0.99
	R	0.97	0.01	0.34	0.97	0.03	0.95	0.83	0.98	0.75	0.77	0.99	0.01
3	L	0.02	0.99	0.85	0.02	0.99	0.03	0.06	0.01	0.74	0.96	0.02	0.98
	R	0.98	0.01	0.15	0.98	0.01	0.97	0.94	0.99	0.26	0.04	0.98	0.02
4	L	0.11	0.98	0.95	0.01	0.99	0.37	0.01	0.01	0.03	0.35	0.00	0.98
	R	0.89	0.02	0.05	0.99	0.01	0.63	0.99	0.99	0.97	0.65	1.00	0.02

The movement away from 0.50 probability is a sort of "rich get richer" effect. When the teacher picks a form and the learner disagrees, the update will move probability toward the teacher's form. Over time, this process will tend to accumulate probability on one of the choices for stress for each of the words.

The consistency across words emerges from the activity of the general constraint. As an illustration, consider the update when the teacher supplies a left-stressed version of Word 1, and the learner's grammar generates a right-stressed one. The difference vector used in the update is shown as $T - L$ in (7).

(7) *A difference vector for a weight update*

		Stress-R	Stress-L	Word-1-R	Word-1-L
Word 1	Teacher: Left		1		1
	Learner: Right	1		1	
	$T - L$	−1	+1	−1	+1

Based on this example of a left-stressed word, the learner will not only raise the weight of Word-1-L and lower the weight of Word-1-R, but it will also raise and lower the weights of the general Stress-L and Stress-R constraints. This increases the probability assigned to left-stressed words in general, and hence increases the probability that when in the future this learner is the teacher it will produce a left-stressed word as a learning datum, leading eventually again to a "rich get richer" snowballing.

To verify that it is the inclusion of the general constraint that leads to the emergence of uniform stress, we ran the simulation without it. This time, uniform stress was only produced in 5 out of the 50 runs. This 0.10 probability comes close to the 0.125 predicted by chance ($p = 0.83$ by a two-sided exact binomial test), unlike the 0.80 probability of consistent stress produced when the general constraint was included.

While the quantitative data remain to be gathered for lexical stress and for other similar cases, we suspect that languages do indeed tend to be regular, as our simulations predict, and as attested regularization in learning and change would lead us to expect. Even though regularity is taken to be the norm in generative phonology, it does not seem that any typological skew in that direction is in fact predicted by other current generative theories. Insofar as a theory incorporates a mechanism to account for morpheme-specific phonology, which is required for descriptive adequacy, then it remains to be explained why languages should tend not to use this mechanism. For some discussion of similar issues in syntax from the perspective of this framework, see Pater (to appear) – the simulation presented here is in fact a slightly disguised version of one presented there for syntactic headedness, with categories, rather than words, providing the basis for the specific constraints. See also Kirby and Hurford (2002: "Why Social Transmission Favors Linguistic Generalization") for useful discussion of why regularization emerges from the dynamics of this sort of system, as well as a comparison with what they take to be standard generative assumptions.[6]

In the next section, we present more elaborate versions of these two simulations, again using MaxEnt phonotactics for the learning simulation, and an OT-style MaxEnt grammar for the typological simulation.[7] The first of these examines the predictions of the learning model for a featural complexity scale originally studied in concept learning. The second of these applies the typological model to making predictions about feature economy.

[6] A methodological advantage of the present Interactive Grammar-Learning model over Kirby and Hurford's Iterated Learning model is that we do not need to find the number of learning trials per generation that will yield what they call the 'bottleneck' effect.

[7] There is nothing that limits the application of each of these grammatical models to each of these domains. For example, we have applied with some success the OT-style model to simulations of learning experiments that involve learning alternations.

2. FURTHER RESULTS

2.1 Predictions for a complexity scale

One reason that structural complexity has been so little studied to date in phonology may be that simplicity biases were assumed to trivially follow from existing grammar and learning models. In both of the preceding sections, we have emphasized that standard generative theories do not make probabilistic predictions about typology, but it may well be that there are many theories in the generative literature that could be realized as probabilistic models with appropriate learning algorithms, and which would generate the same type of basic predictions as we have generated with our models. It is also the case that many cognitive models from outside of linguistics predict simplicity biases, as Culbertson et al. (to appear) have pointed out in a discussion of biases for uniform syntactic headedness. However, it would be a mistake to assume that all simplicity biases trivially follow from any reasonable model. The literature on simplicity biases in concept learning shows that if we make learning problems even slightly more complex than the one in our simulation in 1.1 above, we are able to tease apart the predictions of competing models (see Ashby & Maddox 2005 and Goodman *et al*. 2008 for recent reviews). We expect that this prior research in cognitive psychology will lay the groundwork for much future research in structurally biased phonology. We were first led to this literature by Andrew Cohen's observation that the phonotactic version of IME/CCS presented in 1.1 closely resembles Gluck and Bower's (1988ab) configural cue model, a single layer feedforward network model of category learning that uses as cues each feature, and each of their conjunctions.

The seminal study in concept learning is that of Shepard *et al*. (1961), who examined the relative difficulty of learning categories based on three binary features[8]. The tables in (8) show the 6 ways that a stimulus space can be partitioned into two even-sized categories with three binary features, using the phonological features [+/ labial], [+/–voice] and [+/–continuant]. These 6 Types exhaust the possibilities for formally distinct

[8] The possible relevance of the Shepard hierarchy for phonological learning has been discussed previously by Silverman (1999, 2006) in connection with the structural relationship between allophones of a phoneme. This approach emphasizes the effects of practice in reducing between-type differences in difficulty.

patterns; any other pattern can be obtained by substituting one feature for another. The divisions are indicated by bolding and italicizing the members of one of the categories, which we will arbitrarily call the IN class.

(8) *Shepard et al. patterns as phonological inventories*

Type I		Type II		Type III	
p	***f***	***p***	f	p	f
b	***v***	***b***	v	***b***	***v***
t	s	t	***s***	t	***s***
d	z	d	***z***	d	***z***

Type IV		Type V		Type VI	
p	f	***p***	f	***p***	f
b	***v***	b	***v***	b	***v***
t	***s***	t	***s***	t	***s***
d	***z***	d	***z***	***d***	z

The Type I pattern uses a single feature to divide the space, and here the IN class is [+voice]. The Type II pattern uses two features, in an exclusive-or fashion. Here the IN class consists of segments that are [+labial, –continuant], or [–labial, +continuant]. Types III to V use all three features, but a subset of the stimuli can be described in terms of just two of the features – two of members of the IN class can be identified as the non-labial continuants. The Type VI pattern requires all three features to distinguish each of the segments in the IN class from the OUT segments.

The Saffran and Thiessen (2003) experiment discussed in the introduction examines a pair of patterns that instantiate the Type I *vs*. Type II distinction (though it uses just 6 rather than 8 segments): classes that can be distinguished by voice alone *vs.* ones that need voice and place of articulation. The survey of artificial phonology learning experiments in Moreton and Pater (to appear), which involve a wide range of methodologies with both infant and adult learners, finds several other examples of Type I *vs*. Type II, as well cases of Type II *vs*. Type VI. In every instance, Type I is easier that Type II, and Type II than Type VI.

The results indicate a consistent effect for featural complexity, but they are not particularly informative about the underlying model. For example, all of the models studied in the concept learning literature predict the I < II < VI difficulty ordering.

In discussing the predictions for the full Shepard typology of an IME/CCS implementation of the configural cue model, it will be useful to refer to the abstract characterization of the stimulus space shown by the cubes in Figure 5. Underneath the cubes are the pattern types illustrated with the sort of categories that would be used in a typical concept learning study: objects that are differentiated in terms of binary features of shape, size and color. The IN class is indicated with boxes around the stimuli. The cubes show the categories in a three-dimensional space corresponding to the three features, with the IN class indicated with black dots on the vertices. In these diagrams, the top and bottom faces of the cubes correspond to black and white in the shapes below, left and right to circle and triangle, and front and back to small and big.

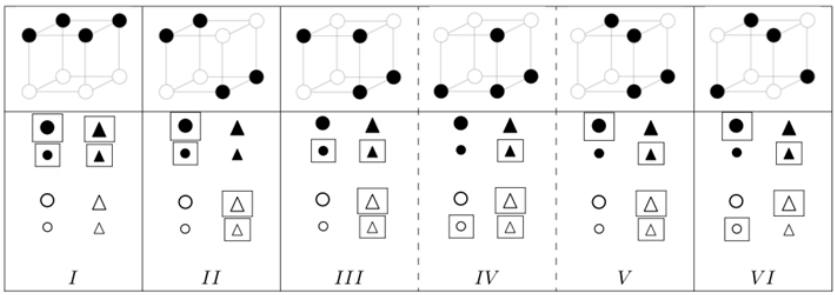

Figure 5. Categories with color, shape and size as features, and cubical abstractions

We implement the phonotactic version of IME/CCS for this sort of space by creating constraints that reward the presence of each feature, and each two-way and three-way conjunction (where here 'feature' means the [–] or [+] value of each binary feature, except that opposite values of the binary features are not combined into a constraint). Each of the 6 single-feature constraints corresponds to one of the faces of the cube, each of the 12 two-feature constraints corresponds to an edge at the boundary of two faces, and each of the 8 three-feature constraints corresponds to one of the vertices at the boundary of three faces.

This way of forming the constraint set is parallel to the simulation in 1.1, except that here we used a fixed constraint set for all of the 'languages' or patterns, whereas the constraint sets for the two patterns in 1.1 were based on the observed features and combinations for each one (this does not affect the outcome in any crucial way). When we learn this model using the Perceptron update rule (or Stochastic Gradient Ascent), we get the following order of difficulty. Like the simulation in 1.1, the learning data consist of a sequence of samples from a distribution over the space of possible forms: the learner saw each of the 4 IN stimuli with equal probability. Predicted greater difficulty means that probability is shifted onto the observed forms more slowly.

(9) *Predicted order of difficulty: phonotactic IMECCS*

Type I < Type III, Type IV < Type II, Type V < Type VI

Why should this order obtain? Note in particular that this is not the order that would be obtained by a model in which difficulty correlates with the number of features needed to define the pattern: as mentioned beneath (8), Types III and IV need more features than Type II. Recall from 1.1 that *ptk* was learned quickly because the [–Vce] constraint gained weight quickly. The *pdk* language had no single-feature constraint that distinguished all of the IN forms from all of the OUT ones. Referring to the cubes now, we can see that for a Type I language, there are two faces that have only IN stimuli (black dots) or OUT stimuli (white dots) at their corners. The first of these corresponds to a single feature constraint that can shift probability onto only IN stimuli when it is given positive weight, and the second corresponds to one that can remove probability from only OUT stimuli when it gets negative weight. There are no such constraints for any of the other Types. Similarly, there are 4 edges in the Type I language that join two corners with IN stimuli, and 4 that join two corners with OUT stimuli. These correspond to the two-feature constraints that target those stimuli to the exclusion of members of the other class, and can thus shift probability onto just IN stimuli, or off just OUT stimuli. No other Type has as many of these. It is the number of such two-feature constraints/edges that distinguish the rest of the Types. Types III and IV have 3 + 3 = 6 each, Types II and V have

2 + 2 = 4, and Type VI has none. When a pattern has more of these constraints that target only IN or only OUT stimuli, it is learned more quickly, and is predicted to be easier.

This order of difficulty is consistent with all of the phonological experiments surveyed in Moreton and Pater (to appear), since it contains the I < II < VI order. It is not, however, completely consistent with the usual full ordering found in supervised learning of visual categories (Shepard *et al*. 1961, 21; Nosofsky *et al*. 1994, 356; Smith *et al*. 2004, 403).

(10) *Attested order of difficulty: supervised concept learning*

Type I < Type II < Type III, Type IV, Type V < Type VI

IME/CCS predicts that Type II should be harder than Types III and IV, but it is usually found to be easier (though cf. Nosofsky & Palmeri 1996 and Love 2002). Because it makes the wrong prediction here, the configural cue model has been rejected in the concept learning literature in favor of alternatives including the revised configural cue model in Nosofsky *et al*. (1994), Kruschke's (1992) ALCOVE, Feldman's (2000, 2006) algebraic model of complexity (cf. Lafond *et al*. 2007), Love *et al*.'s (2004) SUSTAIN, and the Rational Rules model of Goodman *et al.* (2008).[9]

It is thus of no small interest to test the full Shepard typology in phonological learning. Not only will this help to differentiate between models of structural complexity in phonology, but it might also shed light on commonalities and differences between language learning and non-

[9] The Rational Rules model seems particularly amenable to implementation as a model of language learning with a structural bias that would generate distinct predictions from our IME/CCS. This model produces a probability distribution over the derivations of a context-free grammar, which lends it considerable expressive power. Intriguingly, its bias for featurally simple patterns emerges from a prior distribution that is chosen to maximize uncertainty, not simplicity. And usefully, a similar Bayesian model (with a stipulated simplicity prior) has been studied in the context of iterated learning (Griffiths *et al.* 2008).

linguistic category learning.[10] The first such test is a recent study by Moreton and Pertsova (2012) that uses the phonotactic learning paradigm of Moreton (2008). This study replicates once again the I < II < VI order that is common to all models, and finds a partial order on the rest of the Types that does not clearly choose either IMECCS or any of the alternatives from the concept learning literature. Some evidence in favor of the particular model discussed here comes from the behavior of individual types of stimuli across the conditions.

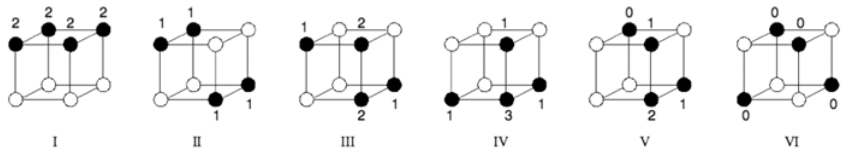

Fig. 6 Number of two-feature constraints rewarding each IN and no OUT stimulus

Figure 6 shows the number of two-feature constraints that reward each of the IN stimuli and that do not reward any OUT stimulus, that is, the number of edges that connect only IN stimuli. The counts for the constraints that reward only OUT stimuli are symmetrical within each Type. Moreton and Pertsova (2012) used a forced choice task that involved choosing between IN and OUT stimuli (CVCV words that either fit a pattern or not – all were nonce words, which had also not been seen in training). They found that across conditions/Types, the probability of choosing the IN stimuli increases with the sum of the number of these "valid edge" constraints that apply to the two stimuli in the trial (which ranges from 0 to 6, since there is a maximum of 3 for the IN and 3 for the OUT stimulus). As Moreton and Pertsova (2012) show, this

[10] We note that there are several differences between the phonological experiments on the one hand, and the psychological experiments in the Shepard tradition on the other, including at least the domain (phonological versus visual), the learning paradigm (usually unsupervised versus usually supervised), the number and kind of irrelevant features, the internal organization of individual stimuli, the ease with which patterns can be described verbally by typical participants, and the perceptual separability of the stimulus dimensions. These differences will have to be controlled or manipulated directly to definitively test any hypothesis that the same, or different, structural factors affect difficulty across domains.

result follows directly from the IME/CCS implementation we have discussed here; it remains to be seen whether it follows from any other model.

2.2 Feature economy

We started this paper with a simple example of feature economy, the tendency for languages to have a voiced consonant at one place of articulation only if they have one at another. In this section, we show that how our model can be extended to make predictions in this domain. To do so, we need to incorporate a mechanism that will yield a preference for contrast. Without such a mechanism, our model will simply prefer systems that use fewer features, and will not yield realistic inventories. Here the functional role of contrast is formalized by building on results in bidirectional OT (see Boersma & Hamann 2008; Boersma 2011 and papers cited there; see Wedel 2004 for a different approach to contrast maintenance). Interestingly, in this model contrast emerges from the dynamics of the system, and is not an independently stipulated principle. Feature economy itself comes from the same effect of the general constraints seen in the earlier simulations.

For this simulation there are 6 words at three places of articulation, with three possible phonetic forms for each word: ones with voiced, voiceless or aspirated versions of the initial consonants. The candidate sets are shown in (11).

(11)

Word 1	[bi]/[pi]/[p^{h}i]	Word 4	[bi]/[pi]/[p^{h}i]
Word 2	[di]/[ti]/[t^{h}i]	Word 5	[di]/[ti]/[t^{h}i]
Word 3	[gi]/[ki]/[k^{h}i]	Word 6	[gi]/[ki]/[k^{h}i]

We will see that the model will tend to produce languages that have laryngeal contrasts at each place of articulation. We take feature economy to be instantiated in the extent to which these contrasts are the same (e.g. all voiced *vs*. voiceless, instead of voiced *vs*. voiceless at one place, and voiceless *vs.* aspirated at another). That is, expect to see featurally economical inventories like that in (12A), where the features that occur at all occur in free combination, rather than inventories like that in (12B), in which some combinations are unpredictably missing.

(12) *Patterns of larngeal contrast across place*

A

	[Lab]	[Cor]	[Dor]
[Voice]	b	d	ɡ
[Plain]	p	t	k
[Aspirated]			

B

	[Lab]	[Cor]	[Dor]
[Voice]	b	d	
[Plain]	p		k
[Aspirated]		t^h	k^h

The constraint set includes three 'Realization' constraints for each Word, demanding each of the three phonetic forms shown in (11). For example,Word-1-[bi] assigns a reward to [bi] when it is a candidate for Word-1. As in (12), we assume three monovalent laryngeal features [Voice], [Plain] and [Aspirated], along with the three monovalent place features [Labial], [Coronal] and [Dorsal]. There were general constraints assigning a reward for each of the laryngeal features, as well as more specific constraints rewarding the co-occurrence of each laryngeal feature with each place of articulation. The most specific constraints are the Realization constraints, which reward a laryngeal-place combination for a given Word.

As in the interactive learning simulation presented in 1.2, a randomly chosen teacher produces a form for learning by the other agent. This time, however, the learner is not supplied with the Word that the phonetic form is associated with. As can be seen in (11), there are two Words that can correspond to each phonetic form. For example, [bi] is a possible phonetic form for both Word 1 and Word 2. A learner would just be given just [bi], and must interpret it as the realization of Word 1 or Word 2.

This is a case of hidden structure learning (Tesar & Smolensky 2000), and we adopt a probabilistic version of Tesar and Smolensky's (2000) Robust Interpretive Parsing to deal with it (as in *e.g.* Boersma & Pater to appear). The decision between underlying Words is made by sampling from the probability distribution defined by the grammar, in particular by the weighting of the Realization constraints that link phonetic form with

the Words. In this use of the grammar Words compete as candidates for a given phonetic form. The candidate Words are just those that have Realization contraints linking them to the given phonetic form – for [bi], these are Word 1 and Word 2. Once this choice is made, the learner then proceeds as usual, generating its own phonetic form for the chosen Word, and then updating the weights.

The constraints started out with zero weights, which resulted in a uniform probability of 0.33 being assigned to each of the three candidates for each Word (and also a uniform probability of 0.5 for each of the two possible Words for each phonetic form in Robust Interpretive Parsing). As in the stress simulation the learning rate was 0.1 and weights were restricted to non-negative values, but here each of the 50 runs had 20,000 trials.

The most basic result is that the learners tended towards a single phonetic form for each word. While the starting probability was uniformly 0.33, in 297 of the 300 Words (6 Words × 50 runs), the average probability assigned to one of the phonetic forms by the learners' final grammars was greater than 0.50, and usually much greater than that. This is similar to what we saw in the stress simulation in 1.2, where the agents came to agree on a single stress position per word.

The learners also seemed to tend to avoid homophony. That is, within each place of articulation, it was more often the case that the two Words (e.g. Word 1 and Word 4) had different choices for the laryngeal feature than would be expected by chance. The chance rate of homophony avoidance would be 6/9 = 0.66, since for each of the two Words there are 3 choices of laryngeal feature, and of the 9 combinations, 3 of them result in the same choice for the two Words. Of the 147 occasions when both Words each had one phonetic form with the majority of the probability, in 125 of these those two phonetic forms were different, yielding a rate of homophony avoidance of 0.85. This intriguing effect awaits statistical confirmation, and a better understanding of the dynamics that produce it. Nonetheless, it appears robust across various types of simulation.

Finally, the choice of laryngeal contrast across places of articulation tended towards uniformity. To get a baseline rate, we start with the observation that at each place of articulation, there are three possible patterns of contrast (aspirated *vs*. plain, plain *vs*. voiced, and voiced *vs*. aspirated). Combining these, there are $3^3 = 27$ possible patterns of contrast. Of these, only 3 use a single laryngeal contrast across places of articulation (0.11), while 18 use two laryngeal contrasts (0.67), and 6 use different laryngeal contrasts at each place of articulation (0.22). We compare this baseline rate to that in the 32 runs that had contrasts at each place of articulation, that is, in which the two Words at a place of articulation each had one candidate that got more than 50% of the probability, and the two such candidates were different. Of these 32, 13 (0.41) had the same pattern of contrast across places of articulation, 16 (0.50) had a distinct pattern of contrast at one place of articulation, and only 3 (0.09) had a distinct pattern at all three places of articulation. Thus, there is a skew towards uniformity of laryngeal contrast across places of articulation: the observed rate of uniform laryngeal contrast (0.41) is much higher than that expected by chance (0.11) ($p < 0.001$ by a two-sided exact binomial test). This effect is arising from the same source as the other simulations in the paper: the constraint set and learning algorithm skews the typology toward what would be described as featurally simple systems. The crucial assumption about the constraint set here is that it contains the general constraints assigning rewards to each of [Voice], [Plain] and [Aspirated]. These are the constraints whose high weight produces generalization across place. If we ran the simulation without them, and had only the specific Voice-Place two-feature constraints, we would not get this effect.

Like the other simulations in this paper, this is only an initial exploration of this domain, but like the others it suggests that this model may yield an account of data that escape current models of phonology. While feature economy may provide evidence for features (Clements 2003, 2009; though see Mackie & Mielke 2011), it is not fully explained by their existence. Inventories exist of varying degrees of economy, and even relatively non-economical ones need to be represented in a theory of phonology, and must be able to be learned. A full explanation also requires something like ease of learning for economical systems (as suggested

by Martinet 1968:483), and a mechanism by which this ease can affect typology. This is what we have formalized here. One important direction for further research is to better understand why feature economy seems to apply with greater strength along particular featural dimensions (see Hall 2011 for relevant discussion), both in terms of the present model, and more generally. This may provide evidence for a richer representational structure over which our constraints are stated, or it may prove to motivate a different model entirely.

3. CONCLUSIONS

We have shown how IME/CCS predicts that patterns described as structurally complex will be learned more slowly, and also how its use in the Interactive Grammar-Learning model of typology generates skews toward structural simplicity. We emphasize that we have used simplicity and complexity as purely descriptive terms throughout. We have not proposed a theory in which the relative complexity of linguistic systems is explicitly measured. Instead, these effects emerge from the structure of the constraint sets used in the grammatical model and from the mechanism for updating constraint weights. We offer this as an observation about our model, rather than as a pre-emptive argument for our approach over any future competitors that might use complexity measures. Since structurally biased phonology has been so little studied, it is important to explore the predictions of a range of formalizations. Nonetheless, we do find the success of IME/CCS to date encouraging, and we see a wide range of further applications.

IME/CCS straightforwardly captures the general finding in the artificial phonology learning literature that featurally complex patterns are learned with more difficulty (section 1.1), and some of its quite fine-grained predictions have already found support in the Moreton and Pertsova (2012) experiment discussed in section 2.1. The systematic study of the role of structural complexity in phonological learning is only just beginning, and there are a number of open research questions. For example: How does supervised learning (≈ alternations) differ from unsupervised learning (≈ phonotactics)? Is the effect of complexity constant across ages of

learners? If not, what is the underlying mechanism that changes? What is the relationship between the effects of complexity in linguistic and non-linguistic category learning? (See further Moreton & Pater to appear).

When implemented in our Interactive Grammar-Learning model of typology, IME/CCS also straightforwardly generates typological skews toward regular stress (section 1.2), and toward feature economy (section 2.2). This opens up further large domains for future research. The empirical data need to be more carefully studied from the perspective of structural complexity, which will require amongst other things better quantification (see e.g. Moreton & Pertsova 2012 on the Shepard types in Mielke's 2008 P-Base). The representational systems of the phonological models will undoubtedly also need to be enriched compared to the simple assumptions we have made here. The nature of the required enrichment will not only depend on further experimental and typological work, but also on a better understanding of how the phonological models interact with theories of phonetics, and with theories of phonological processing (e.g. production and lexical access). It is worth noting that such integration may of course lead to discovering that phenomena that we have attributed to phonological grammar are better explained in terms of phonetics and/or processing. And finally, there is no reason to limit this model of structurally biased phonology to phonology: with the right constraint sets, it also predicts simplicity skews in morphological and syntactic learning and typology.

REFERENCES

Anderson, S. R. 1981. Why phonology isn't 'natural'. *Linguistic Inquiry* 12, 493-539.

Anderson, S. R. 1985. *Phonology in the twentieth century: theories of rules and theories of representations*. Chicago: University of Chicago Press.

Archangeli, D. & Pulleyblank, D. 1994. *Grounded phonology*. Cambridge, MA: MIT Press.

Ashby, F. G. & Maddox, W. T. 2005. Human category learning. *Annual Review of Psychology* 56,149-78.

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

Bane, M. & Riggle, J. 2008. Three correlates of the typological frequency of quantity-insensitive stress systems. In *Proceedings of the tenth workshop of the association for computational linguistics' special interest group in morphology and phonology*: Rutgers Optimality Archive #966.

Berent, I., Wilson, C., Marcus, G., & Bemis, D. 2012. On the role of variables in phonology: remarks on Hayes and Wilson. *Linguistic Inquiry* 43, 97-119.

Blevins, J. 2004. *Evolutionary phonology: the emergence of sound patterns*. Cambridge: Cambridge University Press.

Boersma, P. 1997. Functional optimality theory. In *Proceedings of the institute of phonetic sciences* 21, (pp. 37–42). Amsterdam: University of Amsterdam, Institute of Phonetic Sciences.

Boersma, P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz & J. Mattausch (Eds.), *Bidirectional optimality theory* (pp. 33-72). Amsterdam: John Benjamins.

Boersma, P. & Hamann, S. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25, 217-270.

Boersma, P. & Pater, J. To appear. Convergence properties of a gradual learning algorithm for harmonic grammar. In J. McCarthy & J. Pater (Eds.), *Harmonic grammar and harmonic serialism*. London: Equinox Press.

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

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

Clements, G. N. 1985. The geometry of phonological features. *Phonology Yearbook* 2, 225-252.

Clements, G. N. 2009. The role of features in phonological inventories. In E. Raimy & C. Cairns (Eds.) *Contemporary views on architectecture and representation in phonological theory*. (pp. 19-68). Cambridge: MIT Press.

Coetzee, A. 2002. Between-language frequency effects in phonological theory. Ms., University of Massachusetts Amherst. [http://www-personal.umich.edu/~coetzee/].

Coetzee, A. & Pater. J. 2011. The place of variation in phonological theory. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The handbook of phonological theory* (2nd ed.). (pp. 401-413). Oxford: Blackwell.

Culbertson, J., Smolensky P. & Legendre, G. To appear. Learning biases predict a word order universal. In *Cognition*.

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.

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

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

Goldwater, S. & Johnson, M. 2003. Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, & O. Dahl (Eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory* (pp. 111–120). Stockholm: Stockholm University.

Goodman, N. D., Tenenbaum, J. B. , Feldman, J. & Griffiths, T. L. 2008. A rational analysis of rule-based concept learning. *Cognitive Science*. 32, 108-154.

Griffiths, T. L., Kalish, M. L. & Lewandowsky, S. 2008. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society,* 363, 3503-3514.

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

Hale, M. & Reiss, C. A. 2008. The phonological enterprise. Oxford: Oxford University Press.

Hall, D.C. 2011. Labial place in phonology: universal and variable. Paper presented at NELS 42.

Hannson. G. 2007. On the evolution of consonant harmony: the case of secondary articulation agreement. *Phonology* 24, 77-120.

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

Hayes, B., Kirchner, R., & Steriade, D. (Eds.). 2004. *Phonetically based phonology*. Cambridge: Cambridge University Press.

Hayes, B. & Wilson, C. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379-440.

Hayes, B., Zuraw, K., Siptar, P., & Londe, Z. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85, 822-863.

Hochberg, J. 1988. Learning Spanish stress: developmental and theoretical perspectives. *Language* 64, 683–706.

Jäger, G. 2007. Maximum entropy models and stochastic optimality theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, & A. Zaenen (Eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan* (pp. 467-479). Stanford, California: CSLI Publications.

Jarosz, G. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language* 37, 565-606

Jesney, K. & Tessier, A.-M. 2011. Biases in harmonic grammar: the road to restrictive learning. *Natural Language and Linguistic Theory* 29, 251-290.

Johnson, M. 2007. A gentle introduction to Maximum Entropy Models and their friends. Talk given at the Northeastern computational phonology meeting, University of Massachusetts Amherst. [www.cog.brown.edu/~mj/Talks.htm]

Kingston, J. & Beckman, M. E. eds. 1990. *Papers in laboratory phonology 1: between the grammar and physics of speech*. Cambridge: Cambridge University Press.

Kingston, K. & de Lacy, P. To appear. Synchronic explanation. In *Natural Language and Linguistic Theory*.

Kiparsky, P. 2006. The amphichronic program *vs*. evolutionary phonology. *Theoretical Linguistics* 32, 217-236.

Kirby, S. & Hurford, J. 2002. The emergence of linguistic structure: an overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121-148). Springer Verlag, London.

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

Lafond, D., Lacouture, Y.& Mineau, G. 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.

Liberman, M. 2002. Simple models for emergence of a shared vocabulary. Paper presented at LabPhon 8, New Haven. [http://languagelog.ldc.upenn.edu/myl/labphon.pdf]

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

Madiesson, I. & Precoda, K. 1992. The UPSID database. UCLA.

Martin, A. 2011. Grammars leak: modeling how phonotactic generalizations interact within the grammar. *Language* 87, 751-770.

Martinet, A. 1968. Phonetics and linguistic evolution. In B. Malmberg (Ed.), *Manual of phonetics,* (pp. 464-487) Revised and extended edition. Amsterdam: North- Holland Publishing Co.

McCarthy, J. J. 1988. Feature geometry and dependency: a review. *Phonetica* 45, 84-108.

Mackie, S. & Mielke, J. 2011. Feature economy in natural, random, and synthetic inventories. In G. N. Clements & R. Ridouane (Eds.), *Where do phonological features come from? Cognitive, physical, and developmental bases of distinctive speech categories* (pp. 43–63). John Benjamins.

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

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

Moreton, E. 2008. Analytic bias and phonological typology. *Phonology* 25, 83-127.

Moreton, E. 2010. Constraint induction and simplicity bias in phonological learning. Paper presented at the Workshop on Grammar Induction, Cornell University.
[http://www.unc.edu/~moreton/Papers/MoretonCornellWkshp2010.pdf]

Moreton, E. 2012. Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language* 67 (1):165-183.

Moreton, E. & Pater, J. To appear. Structure and substance in artificial-phonology learning. In *Language and Linguistic Compass*.
[http://www.unc.edu/~moreton/Papers/MoretonPater.Draft.4.1.pdf]

Moreton, E. & Pertsova, K. 2012. Is phonological learning special? Handout of talk given at the 48th Annual Meeting of the Chicago Linguistic Society, with extra section on modeling.
[http://www.unc.edu/~moreton/Papers/MoretonPertsovaCLS2012HO-Big.pdf]

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

Ohala, J. J. 1990. The phonetics and phonology of aspects of assimilation. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I: between the grammar and the physics of speech* (pp. 258-275). Cambridge: Cambridge University Press.

Pater, J. 2000. Nonuniformity in English stress: the role of ranked and lexically specific constraints. *Phonology* 17, 237-274.

Pater, J. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39, 334-345.

Pater, J. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33, 999-1035.

Pater, J. 2010. Morpheme-specific phonology: constraint indexation and inconsistency resolution. In Steve Parker (Ed.), *Phonological argumentation: essays on evidence and motivation* (pp. 123-154). London: Equinox.

Pater, J. To appear. Emergent systemic simplicty (and complexity). In the *McGill Working Papers in Linguistics*. [http://people.umass.edu/pater/pater-systemic-simplicity-mgwpl-2011.pdf]

Phillips, B. 1984. Word frequency and the actuation of sound change. *Language* 60, 320–342.

Pierrehumbert, J., Beckman, M. & Ladd, D. R. 2000. Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr & G. Docherty (Eds.), *Phonological knowledge: conceptual and empirical issues* (pp. 273-303). Oxford University Press.

Prince, A. & Smolensky, P. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, 2004.

R Development Core Team. 2010. R: a language and environment for statistical computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65, 386–408.

Rumelhart, D. & McClelland, J. 1986. On learning the past tenses of English verbs. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing, Volume II* (pp.216–271). Massachusets; MIT Press.

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

Sagey, E. 1990. *The representation of features in non-linear phonology: the Articulator Node Hierarchy*. Garland: New York.

Samuels, B. 2012. *Phonological architecture: a biolinguistic perspective*. Oxford: Oxford University Press.

Scheer, T. To appear. Crazy rules, regularity and naturalness. In J. Salmons & P. Honeybone (Eds.), *The handbook of historical phonology*. Oxford:Oxford University Press.

Schulz, R., Wyeth, G. & Wiles, J. 2010. Language change across generations for robots using cognitive maps. In H. Fellermann, M. Dörr, M. M. Hanczyc, L. L. Laursen, S. Maurer, D. Merkle, P.-A. Monnard, K. Stoy & S. Rasmussen (Eds.), *Artificial life XII: Proceedings of the twelfth international conference on the Synthesis and Simulation of Living Systems* (pp. 581-588.) Masaachusetts: MIT Press.

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

Silverman, D. 1999. On allophonic relations: phonetic similarity or functional identity. Handout, GLOW 99 Workshop on phonetics in phonology.

Silverman, D. 2006. *A critical introduction to phonology: of sound, mind, and body*. Critical introduction to linguistics. Continuum Books.

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

Smolensky, P. & Legendre, G. 2006. *The harmonic mind: from neural computation to optimality-theoretic grammar*. Massachussetts: MIT Press.

Sonderegger M. & Niyogi, P. To appear. Variation and change in English noun/verb pair stress: Data, dynamical systems models, and their interaction. In A. C. L. Yu (Ed.), *Origins of sound patterns: approaches to phonologization*. Oxford: Oxford University Press.

Stampe, D. 1979. *A dissertation on natural phonology*. New York: Garland.

Staubs, R. In prep. Learning biases and main stress-directionality correlation. Ms, University of Massachusetts Amherst.

Tesar, B. & Smolensky, P. 2000. *Learnability in Optimality Theory*. Massachussetts: MIT Press.

Wedel, A. 2004. Category competition drives contrast maintenance within an exemplar-based production/perception loop. *Proceedings of the Workshop of the ACL Special Interest Group on Computational Phonology (SIGPHON)*. Association for Computational Linguistics.

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

Wilson, C. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* 30, 945–982.

Yu, A. 2012. On measuring phonetic precursor robustness: a response to Moreton 2008. *Phonology* 28, 455-518.

Yu, A. (ed.). To appear. *Origins of sound change: approaches to phonologization*. Oxford: Oxford University Press.

Zuraw, K. 2003. Probability in language change. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 139-176). Cambridge, MA: MIT Press.

Joe Pater
Department of Linguistics
University of Massachusetts
Amherst
MA 01003
pater@linguist.umass.edu

Elliott Moreton
Department of Linguistics
University of North Carolina Chapel Hill
NC 27599-3155
moreton@unc.edu