Phonological concept learning

Elliott Moreton
Department of Linguistics
University of North Carolina, Chapel Hill
moreton@email.unc.edu

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

Katya Pertsova
Department of Linguistics
University of North Carolina, Chapel Hill
pertsova@email.unc.edu

Version of August 18, 2014.

Comments are very much welcome, and may be addressed to the authors.

Abstract

Linguistic and non-linguistic pattern learning have been studied in isolation from each other. This paper argues that analogous inductive problems can arise in both phonological and visual pattern learning, that human learners solve them in analogous ways, and that human performance in both can be captured by the same models.

We test GMECCS, which combines the constraint (i.e., cue) inventory of Gluck and Bower (1988a)'s Configural Cue Model with a Maximum Entropy phonotactic framework (Hayes and Wilson, 2008), against the alternative hypothesis that learners seek featurally-simple algebraic rules. We study the full typology of patterns introduced by Shepard et al. (1961) ("SHJ"), instantiated as both phonotactic patterns and visual analogues, using unsupervised training. The main results in both domains differ from the findings of SHJ and the rule-seeking predictions, but resemble each other and the GMECCS predictions. A third experiment tried supervised training (which can facilitate rule-seeking in visual learning) to elicit rule-seeking phonotactic learning, but cue-based behavior persisted.

These results suggest that similar cue-based cognitive processes are available for phonological and visual concept learning.

Keywords: phonotactic learning, concept learning, implicit learning, inductive bias, complexity, Maximum Entropy, Configural Cue Model

1 Introduction

This paper brings together two lines of research that until now have been pursued completely independently: the study of the learning of phonological patterns, and of visual concepts (though see also Moreton and Pater 2012a,b; Moreton 2012; Moreton and Pertsova 2012; Pater and Moreton 2012). Both of these research traditions aim to uncover the inductive biases that humans bring to learning. As we begin to show in this paper, these biases can be better understood by drawing on results from both bodies of literature, conducting controlled comparisons of learning in the two domains and further developing formal models that can capture observed similarities and differences.

In laboratory studies of artificial phonology learning, participants are exposed to a novel pattern that holds over the training data, and then tested on their knowledge of the pattern. To take a simple example, all of the trained words might be constrained to have initial consonants of a particular type. Types of consonants and vowels are defined in terms of phonological features (see Hayes 2009 for an introduction). One such feature is [+/−voice], which separates voiced consonants like [d] and [g] from voiceless ones like [t] and [k] (voicing is produced through vibration of the vocal cords). Consonants are also classified in terms of place of articulation: [d t] are coronal, articulated with the tip of the tongue against the alveolar ridge, while [k g] are dorsal, articulated with the back of the tongue against the velum. The classification of these four sounds by these two binary features is illustrated in figure 1 (a).

Figure 1: A pair of binary features classifying sounds and objects.

a. *Phonological segments*

	Voice	
Place	[+voice]	[−voice]
Coronal	[d]	[t]
Dorsal	[g]	[k]

b. *Visual objects*

	Shape	
Color	Circle	Triangle
Black	●	▲
White	○	△

If we restrict the initial consonants of words to two of these four sounds, then there are two formal types of pattern that we can create (following the Shepard et al. 1961 typology of visual concepts, introduced in section 2.1 below). In Type I, the permitted sounds can be picked out by a single feature. For example, the single feature [+voice] captures the set [d g], allowing us to characterize a pattern in which [da] and [gi] are allowable words, and [ti] and [ka] are not. In a Type II pattern, both features are needed to pick out the permitted sounds. For example, the set [d k] shares no single feature that separates it from [t g], but it can be defined as sounds that are either [+voice] and coronal, or [−voice] and dorsal. As we discuss in our review of this literature in section 2.2, studies with both infants and adults, using a variety of features, have

found that Type I patterns are easier to learn than Type II. It is worth noting, though, that these studies typically use more than just two features over four sounds.

Visual concept learning studies (also referred to as category learning) similarly train participants on a novel pattern, one that holds over a set of visually presented objects (literature reviewed in section 2.1). The features in these studies are properties of the objects. Figure 1 (b) provides an example parallel to the sounds in using just two binary dimensions: shape (circle or triangle) and color (black or white). As might be expected, Type I visual concepts are also easier to learn than Type II, again with the caveat that the spaces of possible concepts are usually larger than in the current simplified example.

Looked at this way, the basic formal relationship between phonological pattern learning and visual concept learning is obvious, but as far as we know, no prior research has capitalized on this connection. There are several benefits to studying phonological and visual concept learning together.

Research on phonological learning can benefit by investigating potential parallels with the vast range of experimental and computational modeling results on visual concept learning. This paper focuses in particular on two architecturally-defined classes of categorization-learning model, which we will call *cue-based* and *rule-based*.

In both kinds of model, an early stage of processing assigns to each stimulus a description in terms of discrete features corresponding closely to sensory dimensions, e.g., [±large] or [±voiced]. Classification and learning are based on Boolean (logical) formulas involving these dimensions. However, cue- and rule-based models use these formulas in very different ways. The differences are highlighted in the following definitions.

By *cue-based*, we mean a model which (1) makes classification decisions based on a weighted sum of a set of cue detectors, where a cue is a Boolean function of a subset of the stimulus features, (2) learns by gradually adjusting these weights, and (3) may penalize changes in the weights of particular cues. A familiar example from category learning is the Rescorla-Wagner model of classical conditioning, in which the cues are the presence or absence of the individual elements of a compound stimulus, the weights are their associative strengths, and the associability (salience) parameters can make the weights of some cues easier to change than those of others (Rescorla and Wagner, 1972; Miller et al., 1995).

By *rule-based*, we mean a model which (1) makes classification decisions using a rule that is a Boolean function of a subset of the stimulus features, (2) learns by testing a small number of hypotheses (candidate rules) at a time (usually just one), and (3) may penalize hypotheses in proportion to their syntactic complexity when stated in the model's rule formalism (usually, in proportion to the number of features involved). A well-known example is RULEX (Nosofsky et al., 1994), which tries each possible one-feature affirmative rule ("has Feature X") before proceeding to two-feature conjunctive rules ("has Feature X and Feature Y"); others include the Boolean-complexity model of Feldman (2000, 2003, 2006) and the rule subsystem of

COVIS (Ashby et al., 1998, 2011).

These two categories, cue-based and rule-based models, are of course not exhaustive (for recent reviews, see Kruschke 2005; Ashby and Maddox 2005; Kruschke 2008; Goodwin and Johnson-Laird 2013). We have chosen to focus on them for the following reasons. First, since important models based on these ideas have been independently developed in both psychology and linguistics, they are a natural starting point for asking whether learning principles are shared across domains. Second, the rule and cue ideas occur as components in hybrid models and in models based on other approaches (e.g., Ashby et al. 1998; Goodman et al. 2008; Love 2002; Maddox and Ashby 2004; Smith et al. 2012). Third, the models that have been actually proposed in these two groups make a number of qualitatively different predictions as to relative pattern difficulty, whether learning is gradual or abrupt, whether learning is implicit or explicit, and how learning difficulty is affected by factors like stimulus verbalizability, deliberate versus incidental learning, rule-seeking instructions, and other task conditions. It is an important question whether these behavioral differences are a necessary consequence of the architectural differences.

A final reason to focus on rule- and cue-based learners is that, at an abstract level, they are kin to (and hence informative about) models with other architectures. Rule-based learners, for example, are related to decision-bound models (Fific et al., 2011), and cue-based models are related to exemplar models via kernel methods (Jäkel et al., 2009). We show in this paper that any member of a large class of Maximum Entropy cue-based learners is equivalent to a member of a larger, and hitherto unexplored, class of similarity-based learners that are based on the Replicator Equation of evolutionary biology (Karev, 2010). Thus, although this paper focuses on rule- and cue-based learners as defined above, the results may prove to be relevant to much larger classes of model.

As we have just noted, cue- and rule-based models of concept learning share ideas with linguistic grammar formalisms (see especially Prince and Smolensky 1993 and Chomsky and Halle 1968 respectively). The relationship can be very close; for example, the cue-based model that we develop and test in this paper is derived from a constraint-based phonological learning model (Hayes and Wilson, 2008), and is very similar to a well-known cue-based theory of concept learning (Gluck and Bower, 1988a,b). The kinship can also be more distant, as in the case of rule-based concept learning models and rule-based grammatical theories. A fundamental difference between rules in the two domains is that rules in linguistics are typically derivational, in that they map one representation to another, while rules in concept learning describe properties of a single level of representation. An important similarity is that the notion of a bias toward rule simplicity, mentioned above for concept learning models, also figures prominently in linguistics, as in the evaluation procedure of Chomsky and Halle (1968). Further exploration of the relationship between rule-based concept learning and (derivational) rule-based grammatical models is an important direction for future research. Our focus in

this paper is on the specific predictions of the cue-/constraint-based model that we develop, and the general predictions of rule-based concept learning models, in particular those that relate to biases favoring simple rules.

Study of the Shepard et al. (1961) (hereinafter "SHJ") patterns has been largely motivated by the differing predictions which these (and other) model classes make about their relative difficulty. The particular kinds of phonological pattern that we are concerned with here are called phonotactics: the restrictions that hold of well-formed words in a language (in contrast to alternations, which refer to changes in the shape of words or morphemes across phonological contexts). The extant phonotactic learning results examine only a subset of the SHJ types, and could be captured by an extremely wide range of learning models. In Section 3 we introduce GMECCS (Gradual Maximum Entropy with a Conjunctive Constraint Schema), a cue-based general concept-learning model that can be applied to phonotactic learning. As we discuss in that section, GMECCS can be seen as a gradual version of the Maximum Entropy phonotactic learner of Hayes and Wilson (2008). We test its predictions against those of rule-based models using all six Shepard types in unsupervised phonotactic learning (Experiment 1, Section 5) and unsupervised learning of visual analogues (Experiment 2, Section 6). The results of these experiments differ from the classic SHJ difficulty order. They are consistent with GMECCS rather than with rule-based alternatives, except that clear signatures of rule-based learning were found for visual patterns of Type I and one subtype of Type II. Since supervised training has been found to facilitate rule-based learning, we also test phonotactic learning of Types I, II, and IV using a supervised paradigm (Experiment 3, Section 7), but still find no evidence of rule-based phonotactic learning.

The results of these experiments also illustrate the benefits that unifying the study of phonological and visual concept learning can have for our general understanding of concept learning. As we discuss in Section 3, GMMECS can also be seen as a Maximum Entropy version of the Configural Cue Model of Gluck and Bower (1988a,b). The Configural Cue model has been revised or replaced in the visual concept learning literature (see esp. Nosofsky et al. 1994) because it fails to capture the classic $I > II > III, IV, V > VI$ order found by SHJ and many subsequent researchers in visual concept learning. Since our empirical results differ from this order in ways that are consistent with the predictions of the cue-based GMECCS, they argue for a reconsideration of cue-based learning models. In this, our results in phonotactic and visual learning confirm and extend recent findings that the SHJ difficulty order, for all its "classic" status, is surprisingly sensitive to task conditions (Kurtz et al., 2013).

The joint study of phonological and visual concept learning also presents a new opportunity to address a fundamental question in cognitive science: Which aspects of language (if any) are acquired using specialized cognitive processes, and which are acquired using domain-general ones? (See recently Jackendoff and Pinker 2005; Christiansen and Chater 2008; Evans and Levinson 2009; Chomsky 2011; Gallistel 2011; Newport

2011.) Phonological and visual concept learning typically take place under different circumstances and involve stimuli with different properties, so to answer them satisfactorily, we must control task and stimulus factors. Taken as a whole, we interpret our results and those of related studies to indicate that when concept learning is relatively implicit, which is encouraged by both lack of verbalizability of features and unsupervised training, it is captured well by GMECCS and similar cue-based models. This holds of both phonological and visual concept learning. As learning becomes relatively explicit, as it is in the classic SHJ paradigm, it is better captured by other kinds of model, including logical rule models. Evidence for this sort of explicit learning that is well-captured by rule-based models is thus far confined to visual concept learning; further study is needed of linguistic learning in this respect. All of these issues are treated in more depth in Section 8.

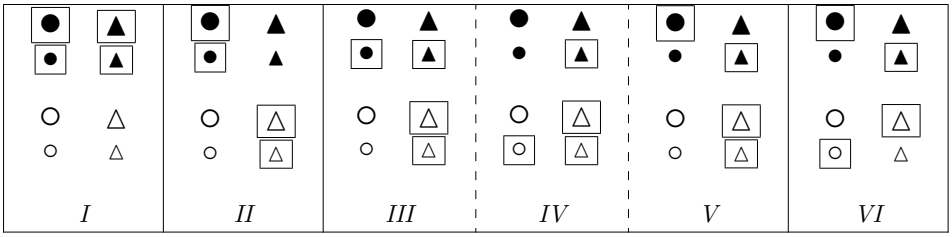
2 Concept learning

2.1 Visual concept learning

Research in visual concept learning has focused on a family of concept types introduced in the seminal study by Shepard et al. (1961), which starts with the observation that a space of concepts defined by three binary features can be divided evenly in six logically distinct ways. (For a larger catalogue of concept families, see Feldman 2000, 2003.) The six Types of concept are illustrated in figure 2, using the three features of shape (circle *vs.* triangle), color (black *vs.* white) and size (large *vs.* small). The members of one of the two classes, which we will arbitrarily refer to as IN, are enclosed in boxes; OUT is unboxed. In Type I, IN can be defined with a single feature value, like black in figure 2. A definition of IN for Type II requires two features, which could be in the form of a logical biconditional − (black *and* circle) *or* (white *and* triangle) − or equivalently, exclusive or: either black or triangle, but not both. Types III - V make crucial use of all three features, but a subset of the members of IN can be defined with just a pair of features: white and triangle in figure 2. Type VI requires all three features to separate any member of IN from all members of OUT.

Shepard et al. (1961) studied learning of each of these types of visual concept by presenting to their experimental participants all eight objects one at a time. After presentation of an object, the participant was asked to classify it as a member of group A or of group B, and was then told the correct classification — this is a form of supervised training. Shepard et al. (1961) found that learning difficulty increased with the number of features needed to define the concept, with Types III - V being easier than Type VI, presumably because a subpart of the space can be defined with less than three features. This $I > II > III, IV, V > VI$ ordering has been replicated in a number of studies that adopt broadly similar tasks and stimuli (see Kurtz

Figure 2: Representatives of the six logically distinct types of equal partition of a stimulus space defined by binary features of color, shape, and size. Boxes enclose one class. After Shepard et al. (1961)



et al. 2013 for a recent review).

The SHJ results have been the focus of intense discussion in the modeling literature. As we will see in detail in Section 3, pure cue-based models have been unsuccessful in matching the classical SHJ order; in particular, they incorrectly predict Type IV to be easier than Type II (Medin and Schwanenflugel, 1981; Gluck and Bower, 1988a). Rule-based models, on the other hand, have been able to replicate the Shepard order, in whole or in part, by using inductive bias that makes them less sensitive to patterns that crucially involve more features (Nosofsky et al., 1994; Feldman, 2000, 2006; Lafond et al., 2007). Selective sensitivity to features, in the form of attentional learning, is also key in the success of the exemplar-based model ALCOVE in modeling the classic SHJ results (Kruschke, 1992).

As Kurtz et al. (2013) emphasize, the classical advantage of Type II over Type IV in human participants can be reduced, eliminated, or reversed by manipulating task and stimulus parameters. When participants were shown only positive category members (unsupervised training), overall proportion correct in Type II fell to or below that in Type IV (Love, 2002). When mention of a rule was omitted from the instructions, the same thing happened (Love, 2002; Love and Markman, 2003; Lewandowsky, 2011; Kurtz et al., 2013). The stimulus dimensions matter as well. When the dimensions were made somewhat harder to verbalize (e.g., speckled vs. checkerboard fill instead of black vs. white fill), overall proportion correct in Type II was reduced relative to that in Type IV (Kurtz et al., 2013, Experiments 2 and 8). The use of perceptually non-separable stimulus dimensions (e.g., brightness, hue, and saturation) also reduced proportion correct in Type II relative to Type IV (Nosofsky and Palmeri, 1996). Even when all dimensions are perceptually separable, some pairs of dimensions are easier to associate in a Type II pattern than others (Love and Markman, 2003; Moreton, 2008, 2012; Kurtz et al., 2013).

One response to the discoveries of a malleable Type II *vs.* Type IV ordering has been to suggest that a rule-based system and a cue-based system compete to control responses, and that task conditions can determine which system dominates (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al.,

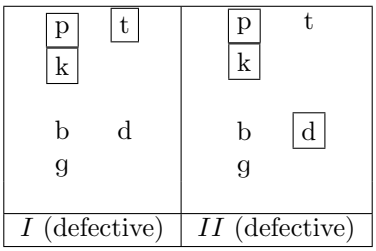
2012). When the rule-based system is dominant, Type II patterns are learned faster than Type IV patterns. When the cue-based system is dominant, Type IV patterns are learned faster than Type II patterns. Much remains to be done, though, in terms of pinning down the empirical effects of task and stimulus variables on the relative difficulty of the Shepard types, and in terms of relating those differences to properties of formal models. For example, when experimental conditions lead to a preference for Type IV over Type II, one possibility could be that the task is primarily tapping a prototype-based learning mechanism that can only represent linearly separable patterns (Medin and Schwanenflugel, 1981). This can be tested by also examining learning on Types III and V, which are not linearly separable.

2.2 Phonological concept learning

Just as visual concept learning studies in the Shepard tradition aim to uncover human inductive biases by comparing the learning of minimally different concepts, so have phonological learning studies aimed to discover whether particular kinds of phonological patterns are easier to learn than others. The general comparison of interest has been between patterns that are (well) attested in the phonologies of the world's languages versus minimally different ones that are not. Many of the specific comparisons have been between patterns that make crucial use of different numbers of features, with the relatively featurally simple patterns taken as more representative of natural language phonological systems. These studies find consistently easier learning for the featurally simpler pattern, using a number of vowel and consonant features, with a wide range of methodologies, and both infant and adult participants. Although previous phonological-learning studies have not intentionally addressed the SHJ hierarchy, a number of studies have independently used phonotactic patterns which can be fitted into it. We describe a few examples here; for a more extensive review, see Moreton and Pater (2012a,b).

Saffran and Thiessen (2003, Exps. 2, 3) found that a pattern that distinguishes [+voiced] [b d g] from [−voiced] [p t k] is learned by English learning 9-month-olds, while a pattern that separates [b t g] from [p d k], which requires both voicing and place of articulation, is not. This is essentially a difference between Shepard Type I and Type II, though it is over a 6, rather than 8 member space (see Figure 3). The training methodology used in this study involved a combination of unsupervised training on pattern induction and word segmentation. In another infant study with training only on positive examples, Cristiá and Seidl (2008) found that English 7-month-olds learned a pattern that separated [-continuant] [m g n t] from [+continuant] [f z], but not [m n f z] from [g t], which cannot be differentiated in terms of any single standard feature. This is also evidence of a Type I pattern being learned more easily than Type II, again with the proviso that the trained concept space had only 6 members (testing involved generalization to a novel sound).

Figure 3: The phonological patterns used by Saffran and Thiessen (2003, Exps. 2 and 3). The features are voiced vs. voiceless, non-coronal vs. coronal, and labial vs. velar.



A Type I advantage was also found using supervised training with adults by LaRiviere et al. (1974, 1977) who trained English speakers with feedback to categorize a set of six or eight syllables into two equal-sized classes. These classes were defined either by a single feature or in an unspecified "random" way that needed more relevant features. In three out of ten experiments, performance was significantly better for the single-feature condition than the random condition, and in the other seven it was numerically better. Evidence for easier learning of phonological Type II than Type VI can be found in a number of adult studies. Kuo (2009) trained Mandarin speaking adults on patterns in which the first member of a word-initial consonant cluster predicted the identity of the second, using exposure to only positive examples. The initial consonants were [t t^h p p^h], which can be distinguished by whether they are aspirated (indicated by superscript [h]) and by place of articulation (coronal [t] *vs.* labial [p]). The second members were the glides [j] and [w], which also differ in place of articulation. This study thus had the full 3 binary features of the Shepard space, but it compared only Type II (e.g. [pj p^hj tw t^hw] as IN) to Type VI (e.g. [pj t^hj tw p^hw] as IN). Kuo (2009) found that two Type II patterns were better learned than Type VI patterns, in terms of novel IN stimuli being significantly more often chosen over OUT stimuli in a forced choice task.

All of the phonological studies we have discussed so far involve learning of *phonotactics*, that is, a set of restrictions on the shape of words. Phonotactic learning studies tend to use unsupervised training, for reasons of ecological validity — outside of the lab, language learners only hear positive examples, i.e., words that really exist in their language. Phonological *alternations* (changes in the pronunciation of individual words or morphemes across contexts) do offer a kind of negative evidence. For example, when an English learner is exposed to the alternating plural morpheme '-s', the fact that it is [s] following (non-strident) voiceless consonants (as in *cats*) indicates that it is not [z] (as in *dogs*) or [əz] (as in *bushes*) in that context. Studies of the learning of alternations have therefore often been run using supervised training, as in Pycha et al. (2003). In that study subjects were asked to judge the well-formedness of words with a suffix in the form of either [-ɛk] or [-ʌk] and were given feedback on their judgement. In the trained patterns, the

choice of suffix was dependent on the features of the preceding vowel. Like the Kuo (2009) study, there were three binary features, used to create Type II and Type VI patterns, and again, performance on two Type II patterns was found to be significantly better than one Type VI. That these findings are robust to changes in methodology is also shown by a study by Skoruppa and Peperkamp (2011) who exposed French speakers to spoken passages in their own language that were modified so that front vowels either agreed in the value of the feature [round] with the preceding vowel (Type II), disagreed, (Type II), or agreed if mid and disagreed if high (Type VI). Participants in the Type II conditions were better at recognizing new pattern-conforming stimuli than those in the Type VI condition.

It is important to note that when patterns are compared that are equivalent in featural complexity, natural language attestedness is not a reliable predictor of ease of learning (Moreton and Pater, 2012a,b).[1] For example, Saffran and Thiessen (2003, Exp. 2) compared two sub-cases of the Type I [+voiced] [b d g] *vs.* [−voiced] [p t k] pattern. In one, the [p t k] set was restricted to the syllable coda (post-vocalic, italicized) position of CVCCVC words, while [b d g] was restricted to onsets (prevocalic, non-italicized Cs). In the other sub-case, it was voiced [b d g] that was limited to coda position, with [p t k] in onset. Although a restriction of [p t k] to coda position is much more common amongst the phonologies of the world's languages (Iverson and Salmons, 2011), both patterns were learned equally well. Similarly, Pycha et al. (2003) and Skoruppa and Peperkamp (2011) examined two sub-cases of their Type II patterns, ones in which two vowels either had the same, or different specification of a feature. Having the same feature, known as vowel harmony, is far more common than vowel disharmony, yet neither study found a statistically significant difference between the two, which is particularly noteworthy in that both studies found a significant difference between the Type II patterns and Type VI. Based on prior results like these, in our experiments we control and test for effects of the featural structure of patterns, but do not control for whether the patterns correspond to ones actually found in natural language or not.

In contrast with visual concept learning, there has been little attempt to computationally model these phonological structural biases (see Wilson 2006 on modeling of a different sort of phonological learning bias, whose empirical basis is questioned in Moreton and Pater 2012b). One reason for this gap in the literature might be that modelling the Type I single-feature advantage over Type II, or Type II over Type IV, has not struck anyone as difficult enough to be interesting. A survey of the learning models proposed for visual concepts shows that this assumption would in fact be correct: all models that we are aware of yield the $I > II > VI$ ordering. In the next section, we show that the $I > II > VI$ ordering emerges from a model

[1] There are two ways of interpreting the mismatches between natural language attestedness and ease of learning of artificial phonologies. One possibility is that the cross-linguistic skews that do not correspond to learning ease have sources in factors other than phonological inductive bias, for example in biases due to articulation and perception, called channel bias by Moreton (2008). The other possibility is that the lab learning studies are not tapping the inductive biases that yield skews in rates of attestation, perhaps because they are not sufficiently ecologically valid.

that has no explicitly stated bias towards patterns with fewer features. In fact, this model does not in general prefer patterns with fewer features, since it also yields $III, IV > II$. (Its predictions thus contradict the SHJ results, but we will see that they are in agreement with the results of other experiments.)

3 The GMECCS model

This section describes GMECCS (Gradual Maximum Entropy with a Conjunctive Constraint Set), a model of concept learning which is applicable to concepts defined phonotactically (Pater and Moreton, 2012). The model combines two main ideas. One is that of a Maximum Entropy (MaxEnt) framework, with training by gradient descent on negative log-likelihood (see Jurafsky and Martin 2008 for an introduction to Maximum Entropy models in natural language processing). This is a cue-weighting model that is closely related to established models in psychology and linguistics: in psychology, the single-layer perceptron (Rosenblatt, 1958), the Rescorla-Wagner model of classical conditioning (Rescorla and Wagner, 1972), and a number of other variations (reviewed in Sutton and Barto 1981); in linguistics, Harmonic Grammar (Smolensky and Legendre, 2006), Optimality Theory (Prince and Smolensky, 1993), and Stochastic Optimality Theory (Boersma, 1998). In applying gradient descent to phonological learning, we follow in particular Jäger (2007), who pointed out the connection to the learning procedure used by Boersma (1998). The other main idea is the unbiased conjunctive constraint set, which provides, as cues available to be weighted, all possible conjunctions of any number of input feature values. Featural conjunctions make up a large share of proposals in the Optimality Theory literature (Ashley et al., 2010). The idea of using them all is adopted from a well-known psychological model of concept learning, the Configural Cue Model of Gluck and Bower (1988a,b). Thus, GMECCS can be viewed as a MaxEnt version of the Configural Cue Model, or as a modified version of the MaxEnt phonotactic learner of Hayes and Wilson (2008). In fact, it combines concepts from both to yield a new model which is equivalent to neither. These relationships are discussed more fully in Sections 3.1 and 3.3 below.

As we discussed in Section 2.1, replication of the SHJ difficulty order has for decades been a major criterion by which models of human category learning were judged. The Configural Cue Model was abandoned because it did not meet this criterion. It has since become clear that human category learning does not always follow the SHJ order (Kurtz et al., 2013), and it is therefore time to reconsider whether the principles underlying the Configural Cue Model may after all correctly characterize learning under certain conditions. Independent developments in phonological theory suggest to us that phonological learning might be one of those conditions. GMECCS is an adaptation of the Configural Cue Model to current phonological learning paradigms.

	+voi	−voi	+voi∧Cor	−voi∧Dor	$h_{\mathbf{w}}(x_j)$	$e^{h_{\mathbf{w}}(x_j)}$	$\Pr(x_j \mid \mathbf{w})$
	$w_1 = -1$	$w_2 = -1$	$w_3 = 4$	$w_4 = 4$			
$x_1 =$ [d]	1		1		3	20.09	0.50
$x_2 =$ [g]	1				−1	0.37	< 0.01
$x_3 =$ [t]		1			−1	0.37	< 0.01
$x_4 =$ [k]		1		1	3	20.09	0.50

Figure 4: A Maximum Entropy grammar defining a probability distribution over a small phonological universe. Constraint weights appear beneath constraint names. The column headed $h_{\mathbf{w}}(x_j)$ shows the weighted sum of scores for each sound, and $\Pr(x_j \mid \mathbf{w})$ the probability assigned to it by the model, which is proportional to $e^{h_{\mathbf{w}}(x_j)}$.

3.1 A Maximum Entropy framework for phonotactic learning

MaxEnt models of phonology use weighted constraints to define a probability distribution over a set of representations (Goldwater and Johnson, 2003; Wilson, 2006; Jäger, 2007; Hayes and Wilson, 2008; Hayes et al., 2009; Coetzee and Pater, 2011). The present model, GMECCS, follows Hayes and Wilson in using MaxEnt to state a probability distribution over the space of possible word forms, thus creating a phonotactic grammar formalizing knowledge of the relative probabilities of word forms (see also Daland et al. 2011; Kager and Pater 2012). It differs from Hayes and Wilson in using a different constraint set, and a different learning algorithm, and in allowing both positive and negative constraint weights. The constraint weights are set using a learner that builds on Jäger (2007)'s application of gradient descent to MaxEnt learning; as we explain below, we use batch, rather than stochastic gradient descent.

The stimulus space (i.e., the universe of phonological words) consists of the set $\{x_1, \ldots, x_n\}$. The model is equipped with constraints $c_1, \ldots, c_m$ that are real-valued functions of x. It is parametrized by a vector $\mathbf{w}$ of weights. The weight parameters control $\Pr(x_j \mid \mathbf{w})$, the model's allocation of probability to Stimulus x_j.

The assignment of probabilities to stimuli is illustrated in Fig. 4 for a very small universe D of phonological representations and a small set of constraints. The constraints in this example target either a single phonological feature, or conjunction of two features. +vce and −vce target voiced and voiceless consonants respectively, +vce∧Cor targets the one consonant that is both voiced and coronal, [d], and −vce∧Dor targets voiceless dorsal [k]. The score in each cell is the number of times the structure targeted by the constraint occurs in the representation (in this example, none happen to exceed 1). For each x_j, this yields a *score vector* $(c_1(x_j), \ldots, c_m(x_j))$.

The *harmony* of x_j is defined as the sum of its score vector, weighted by the current weights: $h_{\mathbf{w}}(x_j) = (c_1(x_j), \ldots, c_m(x_j))^T(w_1, \ldots, w_m)$. The model's estimate of the probability of Stimulus x_j is the exponential of its harmony, divided by the summed exponentials of the harmonies of all representations ($Z_{\mathbf{w}}$):

$$h_{\mathbf{w}}(x_j) = \sum_{i=1}^{m} w_i c_i(x_j) \tag{1}$$

$$Z_{\mathbf{w}} = \sum_{j=1}^{n} \exp h_{\mathbf{w}}(x_j) \tag{2}$$

$$\Pr(X = x_j \mid \mathbf{w}) = \frac{\exp h_{\mathbf{w}}(x_j)}{Z_{\mathbf{w}}} \tag{3}$$

In the example of fig. 4, the weights were deliberately chosen so that the model captures a Type II pattern, dividing nearly all of the probability mass equally between [d], which is both voiced and coronal, and [k], which is neither voiced nor coronal. In practice, the model starts out with all weights equal to zero (causing it to assign equal probability to all d_i), and must learn the pattern-fitting weights from its input $\mathbf{p}$.

The goal of learning is to find the weights that maximize the empirical log-likelihood $L(\mathbf{w})$ of the model (Della Pietra et al., 1997):

$$L(\mathbf{w}) = E_{emp}[\log \Pr(X \mid \mathbf{w})] \tag{4}$$

As our learning algorithm, we adopt the version of gradient descent on negative log-likelihood described by Jäger (2007), in which the weights are changed by an amount proportional to the difference between the true (empirical) probability-weighted average score vector and the learner's current estimate thereof, multiplied by a learning rate η:[2]

$$\Delta w_i = \eta \cdot (E_{emp}[c_i] - E_{\mathbf{w}}[c_i]) \tag{5}$$

Here, $E_{\mathbf{w}}[c_i]$ is the model's current expectation of c_i, obtained by weighting each stimulus's c_i score by the probability currently allocated to that stimulus by the model; i.e., $E_{\mathbf{w}}[c_i] = \sum_{j=1}^{n} p_j \cdot c_i(x_j)$. Likewise, $E_{emp}[c_i]$ is the corresponding sum with the model's probability estimates replaced by the empirical probabilities (i.e., the true probabilities of the stimuli in the distribution from which the training stimuli are sampled).

This model is a unsupervised gradual batch learner. It is unsupervised in the sense that the learning target is a probability distribution over stimuli, rather than an association of stimuli to category labels or other output targets. It is gradual in the sense that the weights (and thus the model's allocation of probability to stimuli) change by a small amount on each update cycle. It is batch in the sense that each step is influenced by the entire target distribution, rather than only by a single sample drawn from that distribution. We chose to use batch rather than on-line learning (i.e., to use gradient descent rather than stochastic gradient descent) because the batch learner conveniently and tractably approximates an average over many independent runs

[2]The following equivalent description of an implementation of gradient descent for MaxEnt may be useful to some readers. A single vector TD representing the training data is produced by first multiplying the vector of constraint scores for each of the representations in the training data by its empirical probability, and then summing over these. A single vector representing the learner's expectation (LE) is generated in the same way, but by using the probabilities generated by the learner's current constraint weights. The difference between these two vectors ($TD - LE$) is multiplied by the learning rate. The resulting vector is then added to the vector of current constraint weights to get the updated constraint weights.

of an on-line simulation.[3] Note that, unlike in some other Maximum Entropy applications in phonology (Goldwater and Johnson, 2003; Wilson, 2006; Hayes and Wilson, 2008), no regularization term is used,[4] and the learning algorithm is gradient descent on likelihood rather than conjugate gradient descent, Improved Iterative Scaling, or any other technique.

We show in Appendix A that it is possible to eliminate the weights from this model entirely.[5] The updates can be done directly on the model's estimated probabilities using the following update rule:

$$\Delta p_j = \eta \cdot p_j \cdot \sum_{i=1}^{n} (c_i(x_j) - q_i) \cdot (q_i^\star - q_i) \tag{6}$$

writing q_i and $q_i^\star$ for $E_{\mathbf{w}}[c_i]$ and $E_{emp}[c_i]$, respectively.

This result is advantageous for three reasons. One is that it allows faster simulation, and thus makes it easier to study the large constraint sets which the conjunctive constraint schema demands. The second, discussed at greater length in Section 4.1, is that the factors of the summand have intuitive interpretations in terms of constraint properties. Finally, as shown in the Appendix, it reveals a connection between GMECCS and the Replicator Equation of evolutionary biology, which has not previously been studied as a basis for psychological learning models.

3.2 Unbiased conjunctive constraints

Because we are interested in how *little* inductive bias is needed to account for the facts, the constraint set in this model is chosen to be as generic as possible, following Gluck and Bower (1988a,b). In a stimulus space defined by N binary features $F_1, \ldots, F_N$, a *conjunctive constraint* $(\alpha_1 F_1, \ldots, \alpha_N F_N)$ is one that targets any stimulus which matches it feature for feature. Each coefficient α can have the values $+$, $-$, or $\pm$ (i.e., "don't care"). A conjunctive constraint of order k on N features is one that has exactly k coefficients that are not $\pm$. (In Fig. 4, the constraint [+voice] is a conjunctive constraint of order 1, while the constraint [+voice]∧[Cor] is a conjunctive constraint of order 2.) Finally, the *complete conjunctive constraint set* on N features is the set of all conjunctive constraints on N features, with orders from 0 to N (a total of 3^N distinct constraints). An example is shown in Table 1. The learners investigated here use only such complete conjunctive constraint sets.

Conjunctive constraints have been previously proposed in the Optimality Theory literature, where they

[3] We have also run on-line stochastic versions of our model, and as would be expected, averaging over runs yields an approximation of the learning curves produced by the non-stochastic version we adopt here.

[4] Regularization is used for several purposes, none of which apply here. One is to prevent weights from going off towards infinity. None of our learning problems have the properties that cause this to happen. Another is to prevent over-fitting. We get generalization from the structure of our constraint sets, and from looking at relatively early stages in the trajectory of gradient descent. Finally, regularization can be used to implement biases in learning, by selectively penalizing particular constraints for departing from a specified value − see especially Wilson (2006). As we discuss in Section 3.2, the inductive biases of GMECCS emerge from the structure of the constraint set.

[5] In fact, as the Appendix shows, it is possible to eliminate them from any model described by these equations, regardless of the constraint set used.

Constraint	Order	Boolean function
$(\pm F_1, \pm F_2)$	0	Always 1
$(+F_1, \pm F_2)$	1	1 iff stimulus is $+F_1$
$(-F_1, \pm F_2)$	1	1 iff stimulus is $-F_1$
$(\pm F_1, +F_2)$	1	1 iff stimulus is $+F_2$
$(\pm F_1, -F_2)$	1	1 iff stimulus is $-F_2$
$(+F_1, +F_2)$	2	1 iff stimulus is simultaneously $+F_1$ and $+F_2$
$(-F_1, +F_2)$	2	1 iff stimulus is simultaneously $-F_1$ and $+F_2$
$(+F_1, -F_2)$	2	1 iff stimulus is simultaneously $+F_1$ and $-F_2$
$(-F_1, +F_2)$	2	1 iff stimulus is simultaneously $-F_1$ and $-F_2$

Table 1: The complete conjunctive constraint set of order 2 on 2 features.

make up a large proportion of proposed feature-based constraints (see the database collected by Ashley et al. 2010). Where this model differs from previous proposals is that the GMECCS constraint set is *unbiased*: it includes all of them, making no attempt to exclude constraints that are implausible for phonetic or typological reasons (in a model of constraint induction, see Hayes 1999) or are unsupported by the training data (Hayes and Wilson, 2008).

3.3 Relationship to previous domain-general learning models

We have just introduced GMECCS as an extension of previous work in modeling of phonological knowledge and learning. In psychology, the MaxEnt framework is closely related to the single-layer perceptron introduced by Rosenblatt (1958). The perceptron calculates a weighted sum of activations over a set of input units that respond to the stimulus. Some versions of the perceptron simply output the sum; in others, the summed activation is compared to a threshold value to yield a binary output. The weights are incrementally adjusted by gradient descent or stochastic gradient descent to minimize the mean squared difference between the output and the desired output. This update rule is called the "Delta Rule" in the case where there is no thresholding (Mitchell, 1997, 94). If an unthresholded perceptron is given only positive category members, and is trained in batch mode to output the value 1 to all of them, then the Delta Rule is identical to the MaxEnt update rule (Equation 5), as long as one is willing to interpret the output activation as the model's estimate of the probability of the stimulus (not always a safe interpretation, since the activation value may go outside of the interval [0, 1]).

The Rescorla-Wagner model of classical conditioning (Rescorla and Wagner, 1972) adds parameters for differences in featural salience, and allows different learning rates on reinforced vs. unreinforced trials, but is still recognizably a perceptron, trained by the Delta Rule (Sutton and Barto, 1981). Gluck and Bower's Configural Cue Model of visual category learning (Gluck and Bower, 1988a,b) consists of an unthresholded perceptron whose input units are, in our terminology, a complete conjunctive constraint set on three features.

GMECCS is simply the Configural Cue Model in the Maximum Entropy framework.

4 GMECCS and the SHJ patterns

The original Configural Cue Model was rejected as a model of human category learning because it did not match human learning performance. Very early in learning, it predicts the order $I > III, IV > II, V > VI$; in a brief middle stage, $I > III > IV > II > V > VI$; then, late in training, $I > III > II > IV > V > VI$ — none of which is the $I > II > III, IV, V > VI$ order found by SHJ (Gluck and Bower, 1988a). There are, we think, two main reasons that justify reviving the Configural Cue Model in the form of GMECCS. One reason, discussed above in Section 2.1, is that the human facts are different from what was previously thought: More recent research shows that the $I > II > III, IV, V > VI$ order is not immutable in visual learning (Nosofsky and Palmeri, 1996; Love, 2002; Kurtz et al., 2013; Crump et al., 2013). The second reason, to be discussed in this section, is that the predictions of GMECCS are different from those of the Configural Cue Model: The obtainable orders are $I > IV > III > V > II > VI$ (early), $I > IV > III > II > V > VI$ (briefly), and $I > III > IV > II > V > VI$ (thereafter). This will first be demonstrated in a 3-feature stimulus space like the one used in all previous experiments on the SHJ hierarchy (Section 4.1). We will then show that the same behavior emerges in an 8-feature space like the one used in our own experiments (Section (4.2).

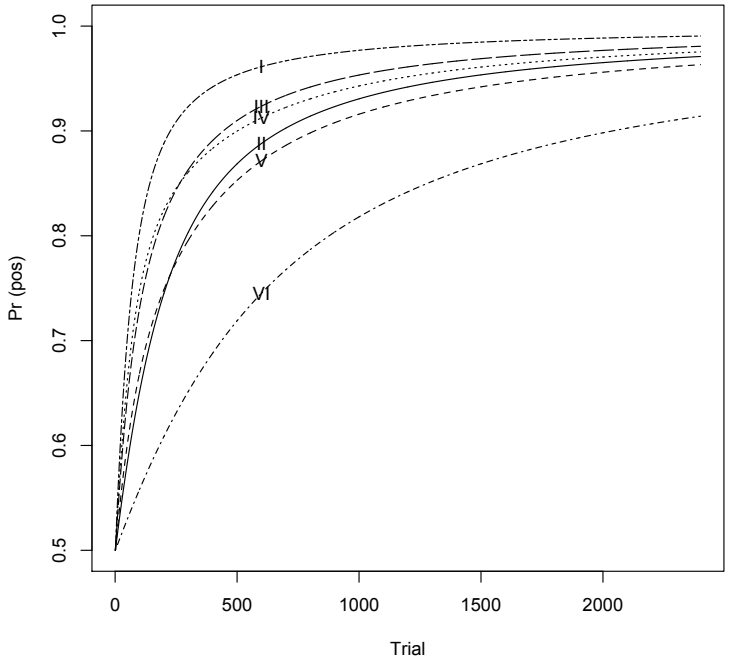
4.1 Learning the SHJ patterns in a three-dimensional space

In this section, we first present GMECCS simulation results to show that the model learns the SHJ pattern types at different rates. Then, we dissect the model's performance to show how these differences arise out of the interaction between Equation 6, the GMECCS constraint set, and the arrangement of individual stimuli within each SHJ Type.

4.1.1 GMECCS simulation

Figure 5 shows GMECCS learning curves for all six types. Each epoch is a single update using the entire empirical probability distribution as described in Section 3; there is no sampling and therefore no variance (two successive runs of the simulation produce identical results). The vertical axis shows the summed probability assigned by the model to all of the positive stimuli together. Constraint weights started at zero, which results in equal probability (1/8) for all stimuli, and a summed probability of 1/2 for all positive stimuli together. GMECCS shifts probability mass onto the positive stimuli fastest for Type I, and slowest for Type VI. Types III and IV are learned faster than Types II and V. Within each these pairs, there is an

Figure 5: Learning curves for Shepard Types predicted by GMECCS in a space of three binary features. The vertical axis shows total probability mass allocated by the model to all of the positive stimuli together. Because the simulations are deterministic, there is no variance.



early advantage for Types IV and V, and a later advantage for Types III and II. The overall $I > III, IV > II, V > VI$ ordering matches the order that Gluck and Bower (1988a, p. 166) report for early learning. However, Type II never overtakes Type IV in GMECCS the way it does in the Configural Cue Model.

4.1.2 Analysis of the simulation

We now ask what is it about GMECCS that causes faster learning for some SHJ types than others. The analysis will focus on the early stage of learning (before the III/IV and II/V reversals) for three reasons. First, the main between-Type differences are present from the very outset. Second, the best quantitative GMECCS fit to our human experimental data occurs at this stage. Finally, the later stages are mathematically less tractable, and we do not have analytic solutions for all Types.

The goal of the early-stage analysis is to understand the source of the differences in the initial slopes of the curves in Figure 5. To preview the conclusion: A positive (negative) stimulus gains (loses) probability quickly if it shares many constraints with other positive (negative) stimuli, since shared constraints allow the model's experience with one stimulus to affect its beliefs about others. Because the GMECCS constraints

correspond to faces, edges, and corners of the SHJ stimulus cubes (see Figure 6), changing the Type changes the pattern of constraint-sharing, and hence the speed of learning, in regular ways.

The correspondence between GMECCS constraints and cube elements is illustrated in Figure 6, adapted from Shepard et al. (1961). The cubes show the concepts in a three-dimensional space corresponding to the three features, with the positive (in-concept) 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. For each corner, edge, and face on the cube (Figure 6), the unbiased conjunctive constraint schema provides a constraint. Face, edge, and corner constraints are shown in Table 2. Each of the single-feature constraints corresponds to one of the faces of the cube, each of the two-feature constraints corresponds to an edge at the boundary of two faces, and each of the three-feature constraints to one of the vertices at the junction of three faces. If we define the *order* of a conjunctive constraint to be the number of relevant features, then face, edge, and corner constraints have orders 1, 2, and 3.

Our analysis of the differences in learning speed across the Types makes use of the concept of a *valid positive constraint* — a constraint that targets (assigns a 1) only to positive stimuli. By *valid constraint*, we mean a constraint that gives a 1 only to positive stimuli, or only to negative stimuli.[6] Which constraints are valid thus varies from Type to Type. A valid positive constraint is thus a valid constraint which gives a score of 1 to positive stimuli only. We use the term *partially-valid constraint* to mean a constraint whose value is 1 on more positive than negative stimuli, or on more negative than positive stimuli. Only Type I has a valid positive single-feature constraint, a face that has only black dots at its vertices. All of the Types, except Type VI, have valid positive two-feature constraints, edges that connect two black dots. Types III and IV have three of these, while Types II and V have only two. This greater number of positive two-feature constraints correlates with the faster learning of Types III and IV; we will now explain why.

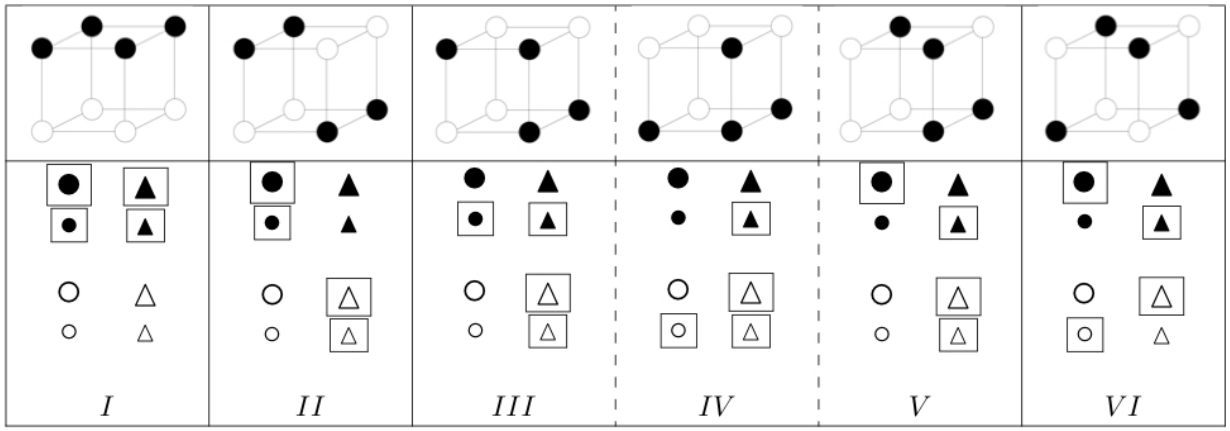
Every stimulus is supported by (receives a score of 1 from) one corner constraint, three edge constraints, and three face constraints, but these constraints can differ in the degree to which they agree or conflict with the categorization of other stimuli. For example, consider the positive stimuli of Type IV. Type IV has a core-periphery structure with one central stimulus — let's call it x_8 — surrounded by three peripheral stimuli, which we can call x_4, x_6, and x_7. The central stimulus is directly supported by its own valid positive corner constraint, by three valid positive edge constraints, and by three partially-valid positive face constraints. An *invalid constraint* is one whose value is 1 on equally many positive and negative stimuli. By contrast, a peripheral stimulus like x_7 is directly supported by one valid corner constraint, one valid edge constraint,

[6] We deviate here from the terminology of Gluck and Bower (1988a), who would call such a constraint *partially* valid unless it is a face constraint.

				Discrepancy		Specificity	
		$q_i^\star$	q_i	$(q_i^\star - q_i)$	$c_i(x_j)$	$(c_i(x_j) - q_i)$	$(q_i^\star - q_i)\cdot (c_i(x_j) - q_i)$
Order 1: Face constraints (one-feature assertions)							
Valid	$\hat{\bullet}\bullet$ $\bullet\bullet$	1	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{4}$
	$\circ\circ$ $\circ\circ$	0	$\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	$\frac{1}{4}$
Part-valid	$\hat{\bullet}\bullet$ $\bullet\circ$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1	$\frac{1}{2}$	$\frac{1}{8}$
	$\hat{\bullet}\circ$ $\circ\circ$	$\frac{1}{4}$	$\frac{1}{2}$	$-\frac{1}{4}$	1	$\frac{1}{2}$	$-\frac{1}{8}$
	$\circ\bullet$ $\circ\circ$	$\frac{1}{4}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$-\frac{1}{2}$	$\frac{1}{8}$
	$\circ\bullet$ $\bullet\bullet$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	$-\frac{1}{2}$	$-\frac{1}{8}$
Invalid	$\hat{\bullet}\bullet$ $\hat{\bullet}\circ$ $\circ\circ$, $\circ\bullet$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{2}$	0
	$\circ\circ$ $\circ\bullet$ $\bullet\bullet$, $\bullet\circ$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$-\frac{1}{2}$	0
Order 2: Edge constraints (conjunctions of two features)							
Valid	$\hat{\bullet}\bullet$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1	$\frac{3}{4}$	$\frac{3}{16}$
	$\circ\circ$	0	$\frac{1}{4}$	$-\frac{1}{4}$	0	$-\frac{1}{4}$	$\frac{1}{16}$
Invalid	$\hat{\bullet}\circ$	$\frac{1}{4}$	$\frac{1}{4}$	0	1	$\frac{3}{4}$	0
	$\circ\bullet$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	$-\frac{1}{4}$	0
Order 3: Corner constraints (conjunctions of three features)							
Valid	$\hat{\bullet}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	1	$\frac{7}{8}$	$\frac{7}{64}$
	$\circ$	0	$\frac{1}{8}$	$-\frac{1}{8}$	0	$-\frac{1}{8}$	$\frac{1}{64}$

Table 2: The conjunctive constraints and their contribution to the model's estimated probability of a positive stimulus (x_j) via Equation 6 at the outset of learning. Each dot array depicts the stimuli that receive a 1 from c_i. Black and white dots $\bullet$ and $\circ$ represent positive and negative stimuli respectively, and x_j is marked as $\hat{\bullet}$. Rotations and reflections are omitted. Columns: $q_i^\star$ is the true expectation of c_i in the target distribution; q_i is the model's current expectation; $c_i(x_j)$ is the value of Constraint c_i on Stimulus x_j. Not shown is the bias constraint, whose value is 1 for all 8 stimuli. A constraint is *valid* if it assigns a 1 only to positive stimuli, or only to negative stimuli.

Figure 6: Shepard Types represented on cubes, after Shepard et al. (1961)



and two partially-valid face constraints.

The rate at which probability mass is moving onto a Type is the sum of the rates at which it is moving onto each positive stimulus, which in turn is determined by Equation 6. Initially, all stimuli are assigned equal probability, so differences between stimuli in the rate of change Δp_j depend only on $\sum_{i=1}^{n}(c_i(x_j)-q_i)\cdot(q_i^\star-q_i)$. Each term of that sum represents Constraint c_i's contribution of new probability mass to Stimulus x_j. Each term is a product of two factors, which we can call *specificity* and *discrepancy*.

Specificity, $(c_i(x_j)-q_i)$, measures how atypical x_j is on constraint c_i, i.e., how specific c_i is to x_j. Among positive GMECCS constraints, specificity is greatest for a corner constraint (which awards a 1 to just one stimulus) and least for a face (which awards a 1 to half of all stimuli). Specific constraints are effective supporters because specificity provides leverage. For example, when the weight of a corner constraint is increased, the increase in the probability mass assigned to that stimulus must be balanced by an equal decrease in the total probability mass assigned to the other seven stimuli; hence, the up-weighted corner gains seven times the probability mass lost by each of the other stimuli.

Discrepancy, $(q_i^\star - q_i)$, measures the difference between the empirical and the estimated probability with which training stimuli (all positive) get a score of 1 from c_i. Each step on the scale valid corner–valid edge–valid face doubles the number of such positive stimuli, and hence doubles the discrepancy. For constraints of a given order, discrepancy is lower for partially-valid constraints because they assign a score of 1 to fewer positive stimuli.

The product of these factors, shown in the last column of Table 2, is the contribution of each kind of constraint, and they are not all the same. If c_i is valid and x_j is positive, then the specificity is $(1 - q_i)$ (because $c_i(x_j) = 1$), and the discrepancy is q_i (because all positive stimuli are twice as probable in the

empirical distribution as the model initially believes they are); hence, each valid c_i's contribution to the sum is $q_i(1-q_i)$. This product is maximized when $q_i = 1/2$, i.e., when c_i is a face constraint, and declines as c_i becomes more specific; thus, positive constraints contribute in the descending order: valid faces, valid edges, and valid corners. Partially-valid faces have lower discrepancy, and fall between valid edges and valid corners.

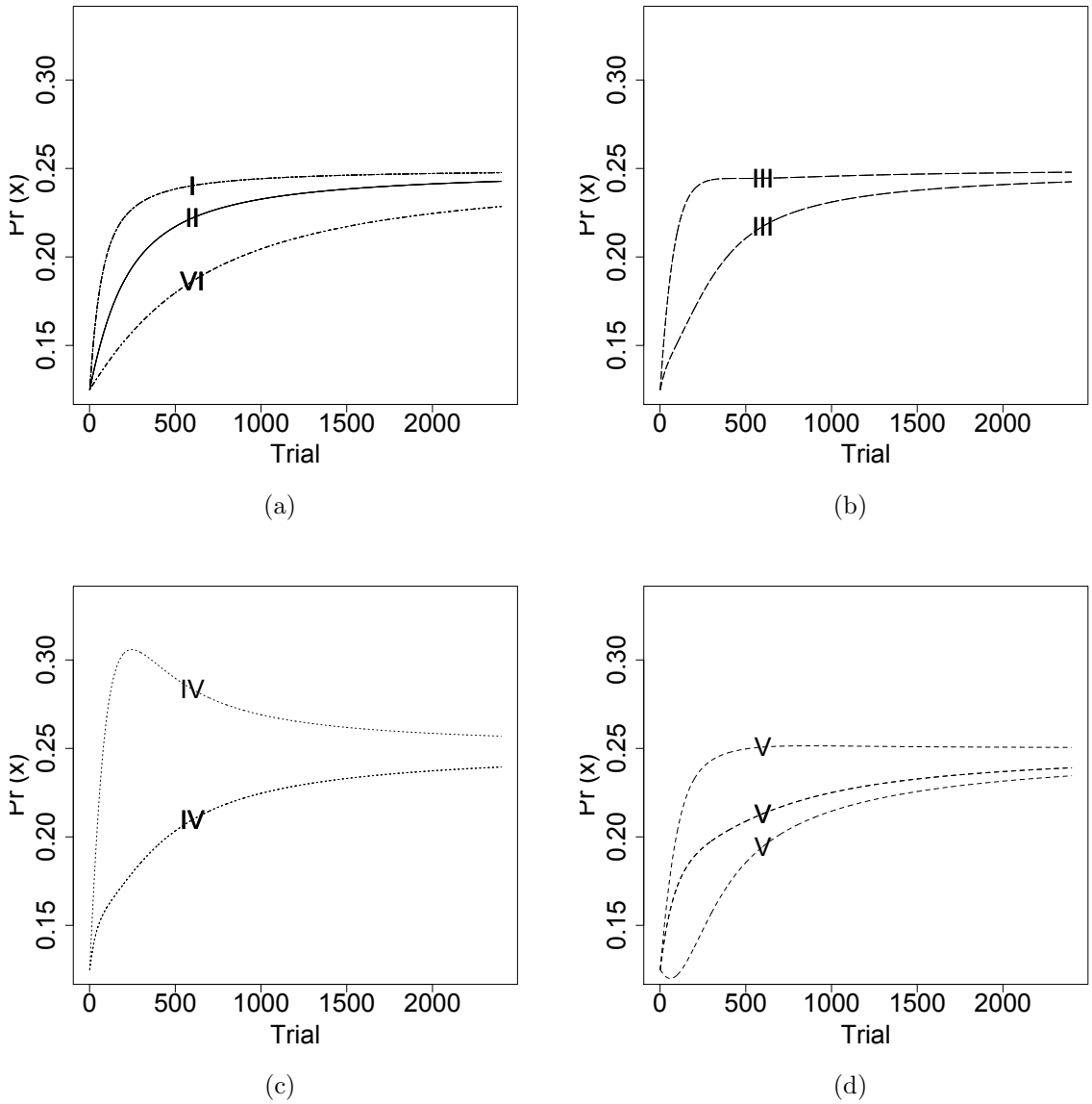
Of these, the most influential across Types are the valid edges. Only Type I affords a valid positive face at all, and even then each positive stimulus is supported by two valid positive edges and only one valid positive face. There are also invalid faces and edges (which contribute nothing). Partially-valid faces may help learning (if three of the four stimuli on that face are positive) or hinder it (if three are negative, because in order to correctly suppress the three negative stimuli, they must also suppress x_j). Negative constraints also support positive stimuli by taking on negative weights and thus suppressing negative stimuli; however, negative edge and corner constraints make small contributions because their specificity for the positive stimulus is low.

From these considerations, we can foresee that, across Types, positive stimuli that are central, in the sense of having two or three positive edge neighbors, will gain probability mass earliest, and those that are isolated (no positive edge neighbors) gain it latest. This is borne out when the simulations are broken down into individual stimuli, as shown in Figure 7; for example, the central stimulus in Type IV, with three edge neighbors, is learned faster than the peripheral stimuli, which have only one.

Since learning performance for a Type is the average across the stimuli in it, performance is determined by the mix of central, peripheral, and isolated stimuli that a Type contains. Type VI is learned most slowly by the model because all four stimuli are isolated. The face and edge constraints are neutral in Type VI, so learning must rely on the corner constraints. Type I is learned quickly because there are no isolated or even peripheral stimuli; face, edge, and corner constraints overlap and reinforce each other. Type II is slow because all stimuli are peripheral. Type V improves on Type II by having a central stimulus, but cancels out the improvement by also having an isolate. Types III and IV are relatively fast because they have no isolates and one or two central stimuli.

This analysis is valid for the initial state, when all of the stimuli are assigned equal probability by the model. It remains a good approximation as long as the probabilities continue to grow or decay exponentially at a rate not too far from their initial values. As learning progresses, the approximation deteriorates, since in reality these rates change over time (e.g., the III/IV reversal occurs because the growth rates in Type IV slow down relative to those in Type III). In the long run, the learner's assignment of probability converges to the empirical distribution, since the corner constraints allow it to match any distribution, and gradient descent is guaranteed to converge.

Figure 7: Probability allocated to each of the positive stimuli within each SHJ Type as a function of time. The learning target is $\Pr(x) = 1/4$. (a) In Types I, II, and VI, all positive stimuli have the same number of positive edge and face neighbors, so there is only one kind of stimulus in each Type. (b) Type III, upper curve: two central stimuli; lower curve: two peripheral stimuli. (c) Type IV, upper curve: one central stimulus; lower curve, three peripheral stimuli. (d) Type V, upper curve: one medial stimulus; middle curve: two peripheral stimuli; bottom curve: one isolated stimulus. (Three-dimensional stimulus space, unsupervised training, learning rate $\eta = 0.01$.)



(a) (b)

(c) (d)

4.2 Generalization across irrelevant features

Phonotactic learning experiments, including our own, typically differ from the classic experiments of Shepard et al. (1961) and their successors (Nosofsky et al., 1994; Nosofsky and Palmeri, 1996; Smith et al., 2004; Lewandowsky, 2011; Kurtz et al., 2013; Crump et al., 2013) in more ways than just using verbal rather than visual stimuli. One major difference is that the stimuli are more complex: Our stimuli are drawn from an 8-feature space rather than a 3-feature space. One, two, or three of the 8 features are used to define the SHJ pattern; the others are irrelevant and vary randomly. Another difference is that our experiments test generalization to new stimuli rather than learning of the training stimuli: Participants are familiarized on 32 of the 128 positive (pattern-conforming) stimuli, and are then asked to choose between a new positive stimulus and a negative stimulus. Finally, participants in our experiments are given no hints as to which of the 8 features are relevant to the pattern, and which ones are irrelevant distractors. Participants must thus do both "attribute identification" and "rule learning", in the terminology of Haygood and Bourne (1965).

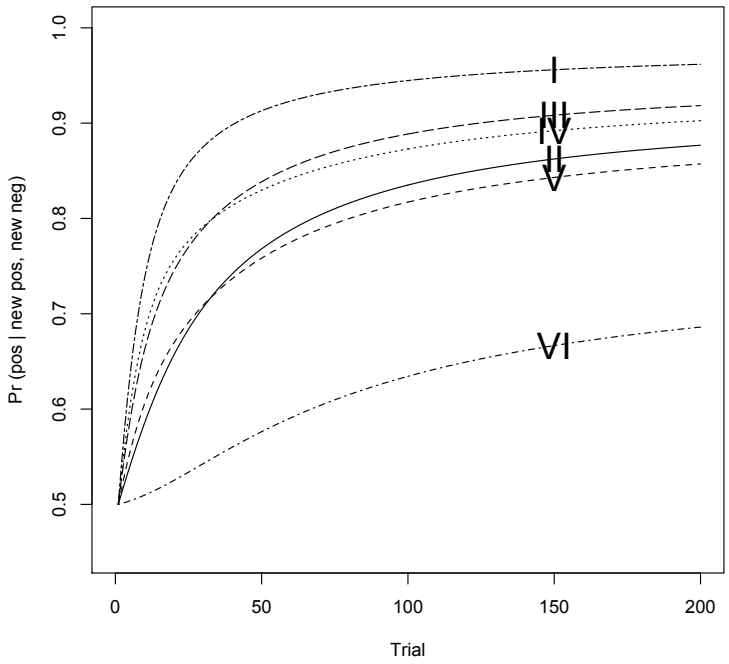
We therefore carried out GMECCS simulations under conditions that were, in the relevant ways, like those of our experiments. The simulations used an 8-feature stimulus space with 256 stimuli, and the constraint set contained all 6561 (3^8) conjunctive constraints of orders 0 through 8. On each run of the simulation (corresponding to one participant in the experiment), 32 training stimuli were randomly selected such that there were 8 training stimuli in each positive cell of the pattern (e.g., for a Type IV pattern, there were 8 training stimuli that belonged to the central cell, and 8 that belonged to each of the peripheral cells). The target (empirical) distribution allocated equal probability of 1/32 to the training stimuli and 0 to all other stimuli, including the other positive stimuli. The learning rate η was 0.01. The two-alternative forced-choice test between a new positive stimulus x_+ and a (new) negative stimulus x_- was modelled using the Luce choice rule (Luce, 1959):

$$\Pr(x_+ \mid x_+, x_-) = \frac{p_+}{p_+ + p_-} \tag{7}$$

The resulting generalization-performance curves, averaged over 250 simulations in each SHJ Type, are shown in Figure 8. They are qualitatively very similar to the learning curves for the three-feature model shown in Figure 5. Performance is always best in Type I. The next best are Types III and IV. Early on, Type IV performance is slightly superior to Type III, but Type III overtakes it later in learning. Likewise, Type V is initially superior to Type II, but falls behind later. Type VI is invariably worst. The possible difficulty orders at different times are $I > IV > III > V > II > VI$, then $I > IV > III > V > II > VI$, and finally $I > III > IV > II > V > V$.[7]

[7]Learning in GMECCS is faster with 8 features than with 3. This effect can be seen in Equation 6: Adding constraints

Figure 8: Predicted generalization performance (predicted probability of choosing a new positive stimulus over a negative stimulus) for 8-feature GMECCS (average of 250 replications per Type). Chance performance is 0.5. Standard deviation for each curve at every time point shown is less than 0.015 for Types I–V, and less than 0.026 for Type VI. Standard error of the mean is standard deviation divided by $\sqrt{250}$. See text for training parameters.



In sum, GMECCS has an inductive bias for learning of Shepard Types with an ordering $I > III, IV > II, V > VI$. This bias emerges from the interaction between the Maximum Entropy learner and the complete unbiased conjunctive constraint set, and is not independently stipulated. The bias is robust against 32-fold enlargement of the stimulus space and 243-fold enlargement of the constraint set, and is present for old (trained) and new (untrained) stimuli alike. Crucially, at no time in learning does GMECCS replicate the original Configural Cue Model's prediction that Type II will overtake Type IV late in learning.[8] GMECCS will therefore be disconfirmed if our experiments replicate the oft-reported advantage for Type II over Type IV (Shepard et al., 1961; Nosofsky et al., 1994; Smith et al., 2004; Minda et al., 2008).

5 Experiment 1: The Shepard patterns in phonological space

There are several reasons to expect that human phonotactic learning should unfold as described by GMECCS, rather than as described by rule-based models of concept learning such as RULEX (Nosofsky et al., 1994). Theoretically, GMECCS is derived from independently-motivated linguistic theories of phonotactics and phonotactic learning (see Section 3), whereas the rule-based models are not. Empirically, cue-based learning of concepts is facilitated by unsupervised training, incidental (unintentional) learning, and stimulus dimensions that are hard to verbalize (Love, 2002; Kurtz et al., 2013) — all properties of the phonotactic learning situation. Experiment 1 is a straightforward test of this hypothesis. It compares the learning of all six SHJ types in an unsupervised phonological-learning paradigm to see whether human generalization performance followed the $I > III, IV > II, V > VI$ prediction derived in Section 4.2.

5.1 Methods

5.1.1 Stimuli

The stimuli were words of the shape $C_1V_1C_2V_2$ where C ranged over [t k d g] and V over [i u æ a]. There were 4 binary phonological features, each of them characteristic of a specific type of segment in the CV-tier (either

increases n without reducing any of the other variables, and so increases the size of the weight update Δp_j. This seems both paradoxical and counter-empirical, since human pattern learning is slowed by irrelevant features (Kepros and Bourne, 1966). The paradox is due to the fact that the learning rate η limits the weight adjustment to each individual constraint, rather than the total weight adjustment to all constraints together. A model with more constraints therefore gets to move further in its hypothesis space on each learning step. The paradox could be removed by scaling the learning rate inversely with the number of dimensions.

[8] The reason for this prediction is that the training target for the Configural Cue Model is *exactly* +1 or −1 on every trial. Consequently, it is possible for the network to be "too good": If the stimulus is a positive one, and the network's output is +2, the update rule will "correct" it downward. In order to correctly classify the three peripheral stimuli in Type IV, the three valid edge constraints have to have high weight. All three of them send activation to the central stimulus, making it "too good" and causing the update rule to down-weigh those edge constraints. That, in turn, reduces performance on the peripheral stimuli. The same logic applies, less drastically, to Type III, where there are two central and two peripheral stimuli. In Types I, II, and VI, all stimuli have the same edge neighborhoods, and there is no such conflict. This allows Type II to overtake the others after long training of the Configural Cue Model. If the Configural Cue Model is trained using what Kruschke (1992) calls a "humble teacher" (one that does not treat "too good" outputs as errors), Type II eventually overtakes Type V, but not Types III or IV.

vowel or a consonant), and each occurring in two distinct syllabic positions (either in the first or the second syllable). This yields a total of 8 binary stimulus distinctions as summarized in Table 3, and a total of 256 possible word types. The stimuli were synthesized using the MBROLA concatenative diphone synthesizer (Dutoit et al., 1996). (These stimuli have been used in several other experiments, including Moreton 2008; Lin 2009; Kapatsinski 2011; Moreton 2012).

	Stimulus position			
	σ_1		σ_2	
Feature	C_1	V_1	C_2	V_2
voiced	±		±	
Coronal	±		±	
high		±		±
back		±		±

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

Table 3: The stimulus space of Experiment 1. The eight stimulus features correspond to the eight non-empty cells. Note that each *phonological* feature is instantiated by two *stimulus* features; e.g., Stimulus Feature #1 is [± voiced], and so is Stimulus Feature #5.

For each participant, three of the eight stimulus features were randomly chosen as the relevant features, and then randomly mapped onto the three logical features defining the Shepard pattern. The 128 pattern-conforming stimulus words made up the "language" for that participant. Thus, each participant was trained on a different "language", i.e., a different instantiation of an SHJ pattern type. Examples are shown in Table 4. Because the patterns were generated at random, they were almost guaranteed to be "crazy rules" (Bach and Harms, 1972; Anderson, 1981), phonetically unmotivated and typologically unattested. This was deliberate, since our aim was to measure purely structural effects on phonological learning. As we noted above, research reviewed in Moreton and Pater (2012a,b), has in any case shown that effects of phonetic motivation and typological attestedness on phonotactic learning are weak compared to those of pattern structure.

Twenty-four patterns ("languages") were generated for each pattern type. Within each of Types II–VI, 6 patterns were generated in each of the cells defined by crossing the two factors *Same Segment* and *Same Genus*. A *Same Segment* pattern was one where two of the relevant features occurred in the same segmental

L1 (TYPE I):	C1 is voiced **d**igu, **g**ada, **d**ika, **g**ugu, ...
L2 (TYPE II):	C1 is voiced iff V2 is back. **d**ig**u**, t**æ**g**i**, **k**ag**æ** **g**ad**a**, ...
L3 (TYPE IV):	At least two of: C1 is voiced, V2 is high, V2 is back **k**ak**u**, **d**ig**u**, **g**uk**i**, **d**æk**a**, ...

Table 4: Examples of artificial phonotactic patterns instantiating SHJ Types I, II, and IV.

position, e.g., "C_1 is voiced iff C_1 is coronal". A *Same Genus* pattern was one where two of the relevant features instantiated the same phonological feature, e.g., "C_1 is voiced iff C_2 is not voiced". Since both factors could not be simultaneously true for Type II, 12 patterns, rather than 6, were generated for the Type II cell where both were false.

The reason for distinguishing the *Same Segment* and *Same Genus* subcases is that GMECCS, like all of the other learning and categorization models discussed in this paper, approaches the learning task without any expectation that particular features will be more closely related to each other than others. This may be true of humans for the shape, size, and color of visual stimuli (but see Kurtz et al. 2013, and below on Experiment 2), but we have grounds to suspect that it may not be true of human phonological learning. If two relevant features of a pattern are in the same segment, a participant could solve (or partly solve) it by analyzing it as a segmental rather than a featural pattern (e.g., for a Type II pattern like "C_1 is voiced iff C_1 is coronal" is equivalent to "C_1 is either [k] or [d].")

From the 128 pattern-conforming words, 32 were randomly chosen as familiarization items, subject to the restriction that 8 words represent each of the four pattern-conforming combinations of relevant feature values. Another 32 were chosen in the same way for use as positive test items, as were 32 non-conforming items which served as negative test items.

5.1.2 Procedure

A participant was seated in front of a computer screen in a sound-proof booth. They were told that they would learn to pronounce words in an artificial language, and later be tested on their ability to recognize words from that language. They listened to and repeated aloud 32 randomly-chosen pattern-conforming stimuli 4 times over. After this training period, they proceeded to the testing phase during which they heard 32 randomly-chosen pairs of new stimuli (one pattern-conforming, one not) and tried to identify the one that was "a word in the language you were studying" by pressing a button on the screen corresponding to "word 1" or "word 2". If a participant gave fewer than 12 correct responses out of 32,[9] their data was discarded and they were replaced. After the experiment, participants filled out a questionnaire in which they were asked whether they noticed any patterns that helped them to complete the task.

5.1.3 Participants

Volunteers were recruited from the UNC-Chapel Hill community using flyers and mass email. They were self-screened for normal hearing and native English. They were paid US$7 for the half-hour experiment. Each participant was randomly assigned to one of the 6 language types with 24 participants per type, for a

[9]This criterion was chosen because it is significantly below chance at the 95% level by an exact binomial test.

Type	I	II	II-sg	II-ss	III	IV	V	VI
Plotting symbol	○	○	△	+	○	○	○	○
Mean	0.737	0,567	0.598	0.708	0.668	0.704	0.651	0.594
SD	0.120	0.112	0.123	0.153	0.129	0.091	0.091	0.099
N	24	12	6	6	24	24	24	24
sem	0.024	0.032	0.050	0.062	0.026	0.018	0.018	0.020

Table 5: Mean proportion correct in each Type condition of Experiment 1. Plotting symbols refer to Fig. 9.

total of 144 participants. An additional 12 volunteers participated, but their data could not be used (6 had non-native English, 1 interrupted the experiment, 1 failed to meet the 12-out-of-32 accuracy criterion, and 4 suffered equipment or experimenter error).

5.2 Predictions

Since participants are tested after a fixed number (128) of training trials, this experiment cannot observe learning curves. Instead, each participant's performance is observed at a single point along their individual learning curve. However, it is clear from the simulations in Section 4.2 that no matter where that point is, GMECCS predicts generalization performance to follow the order $I > III, IV > II, V > VI$. Rule-based models which successfully capture the classic SHJ order predict $I > II > III, IV, V > VI$. The relative difficulty of Types II and IV is therefore critical (Love, 2002; Kurtz et al., 2013).

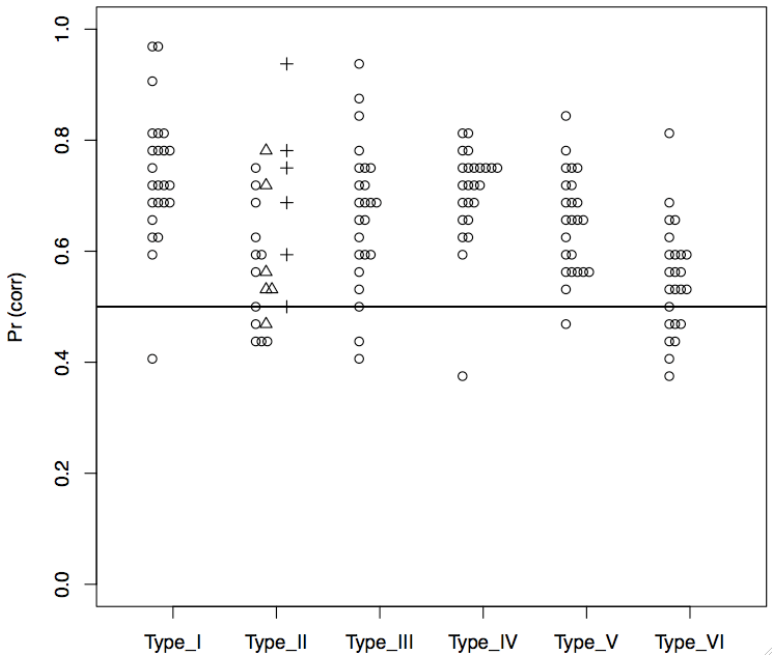
Better performance on *Same Segment* or *Same Genus* patterns than on others would be a sign that participants were influenced by the internal structure of the stimulus, contrary to the assumption of GMECCS, RULEX, the Configural Cue Model — and indeed of all of the models considered in this paper — that stimulus features are interchangeable. Since the use or non-use of internal structure is orthogonal to the question of global model architecture, the relevant test of GMECCS is its fit to the data once the effects of internal structure have been removed.

5.3 Results

Fig 9 shows the mean proportion of correct responses for each subject. Each plotting symbol represents one participant. The vertical axis is proportion of correct (i.e., pattern-conforming) responses. Numerical values are shown in Table 5. The performance means decrease in the order $I > IV > III > V > II > VI$ (excluding the same-segment and same-genus Type II cases). This differs from the classic SHJ order $I > II > III, IV, V > VI$, but matches the $I > III, IV > II, V > VI$ order predicted by GMECCS.

A mixed-effects logistic-regression model was fitted with the `lmer` function of the *lme4* package in R 2.7.1, with a random intercept for each participant. For the *Type* variable, Type II was the reference category, and

Figure 9: Individual participant performance (proportion pattern-conforming responses) for Experiment 1. Plotting symbols for Type II: + = all relevant features are in the same segment analogue (Type II-ss); △ = all relevant features are of the same genus (agreement/disagreement pattern, Type II-sg); ∘ = other.



the other Type-conditions were dummy-coded (e.g., for a participant in the Type VI condition, the factors *I, III, IV*, and *V* were all 0, and *VI* was 1). Preliminary analysis showed that the *Same Genus* and *Same Segment* counterbalancing factors had no effect in Types III—VI, so these sub-conditions were collapsed together within each of these Type conditions.[10] They were retained for Type II, and renamed to indicate their restriction to Type II. A Type II pattern for which *Both Same Genus* = 1 is a harmony or disharmony pattern (e.g., V_1 is [+high] iff V_2 is [−high]), while one for which *Both Same Segment* = 1 has both relevant features occurring in the same segmental position.

Two further factors, *Redup* and *CorrFirst*, were included to absorb the effects of aversion to reduplicated stimuli (e.g., [gigi]), and of preference for the first of the two-alternative forced-choice stimuli. *Redup* was 1 if the pattern-conforming response on a particular trial was reduplicated but the nonconforming response was not, −1 if the reverse, and 0 if both or neither stimulus was reduplicated. *CorrFirst* was 1 if the pattern-conforming response was presented first, else 0 (Moreton, 2008, 2012).

The fitted model is shown in Table 6. The intercept did not differ significantly from zero, indicating that performance in the basic Type II condition was not significantly above chance levels. Significant positive

[10]Since all Type I patterns involve only one feature, it is also possible to view Type I patterns as having all features in the same segment and all features of the same genus. There is no way to analyze our results to determine how much of Type I's advantage comes from using only one feature, and how much comes from not mixing segments or genera.

effects were found for the Type variables I, III, IV, and V, but not for VI; thus, performance was better than Type II (and hence better than chance) in every Type condition except Type VI. Within Type II, *Both Same Genus* did not differ significantly from zero. The significant positive estimate for *Both Same Segment* shows that Type II patterns were much easier — as easy as Type IV, and nearly as easy as Type I — when both of the relevant features occurred in the same segment. Finally, both of the nuisance factors *CorrFirst* and *Redup* had significant effects, indicating that participants were biased to choose the first of the two response options and to avoid reduplicated words.

Coefficient	Estimate	SE	z	$\Pr(>\|z\|)$
(Intercept)	0.12803	0.14337	0.893	0.371865
I	0.81999	0.17583	4.663	< 0.0001
III	0.47202	0.17322	2.725	0.006432
IV	0.63399	0.17423	3.639	0.000274
V	0.38984	0.17265	2.258	0.023952
VI	-0.06134	0.17119	-0.358	0.720114
Both Same Genus	0.11078	0.24329	0.455	0.648850
Both Same Segment	0.69318	0.25149	2.756	0.005845
CorrFirst	0.27396	0.06348	4.316	< 0.0001
Redup	-0.77231	0.10142	-7.615	< 0.0001

Table 6: Summary of fixed effects for the mixed-logit model for Experiment 1 (4608 responses from 144 participants; log-likelihood = -2879).

The statistical analysis thus confirms that, unlike in the classic Shepard experiments, and contrary to the predictions of models which successfully capture the results of those experiments, when other factors are controlled Type II is not easier than Types III, IV, and V, but harder, as predicted by GMECCS. GMECCS has only one free parameter, the learning rate η (in our simulations, always 0.01). Within broad limits, changing η changes only the time scale, and is equivalent to multiplying the trial numbers by a constant factor. To find the best GMECCS fit to the human data, we found it convenient to leave η as it was and instead find the trial number that minimized the disparity between human and GMECCS performance. Human performance was taken to be the cell means in the fitted logistic-regression model of Table 6, where Type II was represented only by the basic Type II subcell.[11] By using the modelled cell means rather than the raw ones, we removed the effects of the nuisance variables (which GMECCS is not designed to account for). Disparity at a given trial was taken to be the squared difference between these cell means and the GMECCS proportion correct at that trial. The best match (disparity < 0.0032) was

[11] As noted in Section 5.1.1, GMECCS, like RULEX, the Rational Rules Model, the Configural Cue Model, and all of the human learning and categorization models considered in this paper, treats stimuli as unstructured sets of feature values. They do not address aspects of human learning behavior for which that assumption is invalid. The *Same Segment* and *SameGenus* factors were therefore designed into the experiment in order to check the validity of this assumption for human learners in this experimental situtaion, and to enable the *statistical* model to model out effects of factors not accounted for in the *learning* models. The very high performance in the single-segment Type II condition is consistent with our prediction (in Section 5.1.1) that human participants might solve a single-segment Type II problem as a segmental rather than featural pattern.

attained on Trial 9 of the GMECCS simulation, as shown in Fig. 10. (To make GMECCS attain the same performance after 128 trials, like the humans, η would be set to $0.01 \times 9/128 = 0.00070$) The relative numerical order, $I > IV > III > V > II > VI$, was in agreement with the GMECCS predictions (Section 5.2). Performance on Type II was lower than the GMECCS fit (by a margin of 0.046), and on Type V higher (by 0.030), but the other four Type conditions were matched very closely (within 0.012). A logistic-regression model with the following coefficients would fit the Trial 9 GMECCS results perfectly: *(Intercept)* = 0.310, $I = 0.644, III = 0.310, IV = 0.395, V = 0.083, VI = -0.277$. Each of these values falls well inside the 95% confidence interval around the corresponding coefficient estimates from human performance in Table 6.

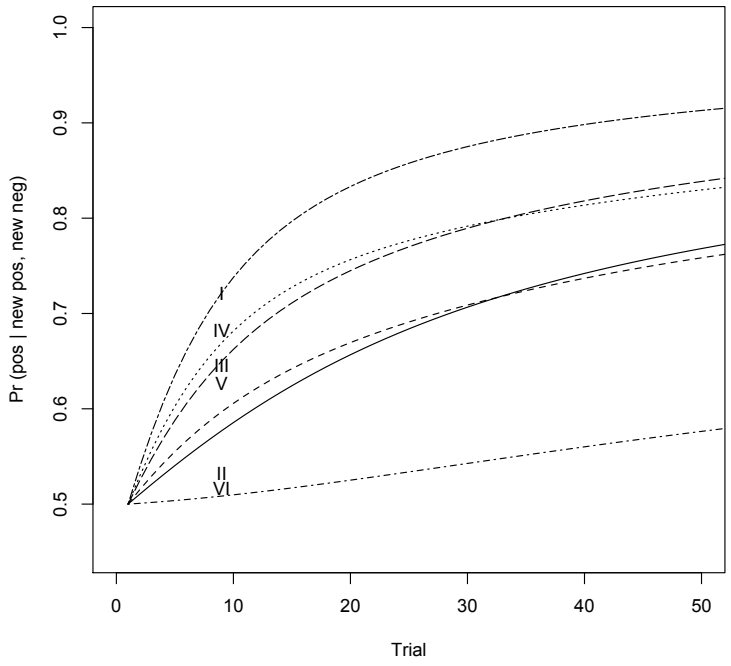


Figure 10: Human performance most closely matches GMECCS predictions at Trial 9. (The Type II mean excludes the same-segment and same-feature-genus cases.) Chance performance is 0.5. Standard deviation for each curve at every time point shown is less than 0.015 for Types I–V, and less than 0.026 for Type VI. Standard error of the mean is standard deviation divided by $\sqrt{250}$. See text for training parameters. The plotting symbols show performance at Trial 9, and also label the curves (V and II are above and below their respective curves).

At this early stage, edge constraints still dominate the behavior of GMECCS (see previous discussion in Section 4.1). The probability that GMECCS will choose a positive stimulus x_+ over a negative stimulus x_- increases with the number of positive edge neighbors of x_+ and the number of negative edge neighbors of x_- (see Section 4.1 above). Figure 11 shows two-alternative forced-choice performance by humans and by

GMECCS on Trial 9 as a function of supporting edge constraints. The highest performance in both cases was attained when there were 6 supporting edge constraints; this occurred on Type IV trials pitting the central positive stimulus against the central negative one. The lowest performance in both (below the 0.5 chance level) occurred when neither stimulus had an edge neighbor, i.e., all Type VI trials and the Type V trials that compared two isolated corners. Between these extremes, performance increases along with edge support for both humans and GMECCS (when edge support is added to the logistic-regression model above, it is highly significant; estimate = 0.360, se = 0.048, $z = 7.421$, $p < 0.00001$).

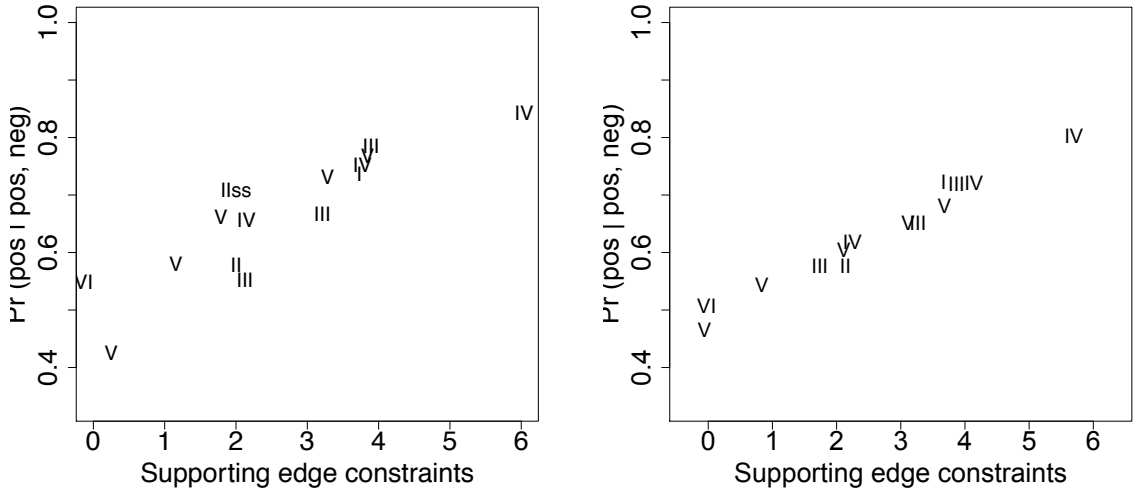


Figure 11: Effect of edge constraints in human learners (left panel) and in GMECCS (right panel) on Trial 9. Horizontal axis: number of positive edge neighbors of positive stimulus plus number of negative edge neighbors of negative stimulus. Vertical axis: Probability of choosing positive stimulus (cell means for actual or simulated data). Plotting symbols have been horizontally dithered for legibility. IIss = Type II with both features in the same segment (humans only).

5.4 Discussion

Where the predictions of GMECCS differed from the classic SHJ order, participants in Experiment 1 followed the GMECCS predictions. The resemblance to GMECCS extended below the level of the SHJ Types when they were broken down further by the edge neighborhoods of the individual stimuli. The results are thus consistent with GMECCS in a detailed way, and support the use of cue-based concept learning and constraint-based phonological models more generally. We now turn to the question of whether they are also consistent with rule-based models, recalling that by "rule-based", we mean a model whose inductive bias respecting a hypothesis is most transparently expressed in terms of the syntactic complexity of the model's representation of that hypothesis (see Section 1). Can the results of Experiment 1 be understood as the result of a preference

for syntactically-simple hypotheses? The answer will depend on the model's syntax as well as on the humans' data.

One proposal is that hypotheses are represented as Boolean disjunctions of conjunctions of individual feature predicates, that hypotheses are penalized as these conjunctions become more numerous or contain more features, and that concept-learning performance depends on the least-penalized hypothesis that correctly describes the pattern (Feldman, 2000, 2006). However, since Types II and III can each be expressed as a disjunction of two two-feature conjunctions, the penalty function, no matter how it is parametrized, does not distinguish them (Lafond et al., 2007), and thus leaves unexplained both the classic SHJ order ($II > III$) and that found in Experiment 1 ($III > II$).

RULEX (Nosofsky et al., 1994) tests rule hypotheses in increasing order of the number of stimulus features involved, trying first one- and then two-dimensional rules until a tolerably accurate one has been found, or until all possibilities have been exhausted, after which it begins memorizing cases not predicted by rule. A perfectly accurate two-dimensional rule exists for Type II, but for Types III–V, the best that can be done is a one-dimensional rule that is 75% accurate. If the model has a low tolerance for inaccurate rules, then Type II will be learned faster than Types III–V, as in Nosofsky et al.'s simulation of their SHJ replication. However, we can imagine a scenario in which the model is content with only partial accuracy. Now the shoe is on the other foot, because there is no one-dimensional rule for Type II that is more than 50% accurate. We might therefore observe the learner at a stage where it has found a partially-accurate rule for Types III–V, but is still testing and discarding candidate rules for Type II that are no better than chance. Could that account for the $III, IV, V > II$ ordering in Experiment 1?

If a 75%-accurate rule is used for a two-alternative forced-choice task, the probability of success is the probability that the rule classifies both stimuli correctly, plus the probability of guessing correctly if the rule mistakenly classifies both stimuli alike, i.e., $(0.75 \times 0.75) + 0.5 \times (0.25 \times 0.75 + 0.75 \times 0.25) = 0.75$. Human performance on Types III–V in Experiment 1 was less, about 0.67, which could reflect error in finding or applying the rule, e.g., if on about 11 trials out of 16, participants were unable to access the rule and had to guess (with 50% success), or if 11 participants out of 16 were unable to find the rule. However, in that case, a 100%-accurate rule, available in the Type I condition, would yield a success rate of at least 0.84, which is more than three standard errors above the performance obtained in the Type I condition of Experiment 1 (Table 5). Thus, in human performance, Type I is not as special as RULEX predicts it to be; instead, the Type I stimuli are treated like any other stimuli that have two edge neighbors, as predicted by GMECCS (Fig. 11).

In the Rational Rules model (Goodman et al., 2008), the hypothesis space is defined by a grammar of logical formulas. The formulas are assigned prior probabilities on the basis of their syntactic structure, and

learning yields a posterior distribution over the whole space. Since Rational Rules monitors a large number of hypotheses simultaneously, but also applies a syntactic prior to them, it falls into neither the "rule-based" nor the "cue-based" group, but occupies a middle ground. Nonetheless, the results of Experiment 1 are informative about it. The Rational Rules model is similar to RULEX in that the order of Type II vs. Types III–V can be manipulated by adjusting a parameter that controls tolerance for rule inaccuracy. Goodman et al. report that predicted Type II performance falls below that for Types III–V when this parameter, b, is reduced to 1. However, the interpretation of this parameter is that the model expects each training datum to be falsely labelled with probability e^{-b}; hence, the reversal is observed only when the model expects 36.7% or more of the training data to be incorrect. This strikes us as unlikely (though not impossible) for human participants in these experiments, or in the ones reported by Kurtz et al. (2013), which used a three-feature stimulus space.

These conclusions must be taken with caution, because none of these models is designed for unsupervised learning, and in no case has their behavior on SHJ patterns been studied for generalization to exemplars outside the training set. However, the principle underlying the models — that inductive bias favors hypotheses that involve fewer features — is not borne out by the present results.

6 Experiment 2: Visual analogues

The results of Experiment 1 diverge from the findings of Shepard et al. (1961) in that Type II proved harder, not easier, than Types III, IV, and V. The starkness of the difference appears to invite the inference that phonotactic patterns are learned using different cognitive resources from non-phonological patterns. Phonotactic learning in particular seems to involve a cue-based process like that described in Section 3, rather than the rule-based processes that would produce the Type II advantage characteristic of the classical experiments.

However, the situation is not that simple. Because Experiment 1 was designed to resemble typical phonotactic-learning experiments (which in turn are designed to resemble natural language learning), it differed from the Shepard experiments in several ways besides just being phonological. Some of those differences are known to affect relative pattern difficulty, especially of Types II and IV, in non-linguistic learning. Experiment 2 asks whether a non-linguistic analogue of Experiment 1 would in fact replicate the classical difficulty order. (We note that phonological and non-linguistic patterns have been compared by earlier researchers, e.g., Smith and Minda 1998; Weinert 2009; Finley and Badecker 2010; Lai 2012.) If the classical order is replicated, that would corroborate the hypothesis that it is phonological learning, in particular, which is served by a cue-based learner of the sort described in Section 3. But if the relative

difficulty of the non-linguistic analogues is similar to that of the corresponding phonotactic patterns, that would be consistent with the hypothesis that, when task conditions are controlled, learning takes place in the same way in both domains.

We are concerned in particular with four differences between the classical experiments on the one hand (Shepard et al., 1961; Nosofsky et al., 1994; Smith et al., 2004) and Experiment 1 on the other which could in principle account for the difference in relative difficulty. First, the classical experiments used supervised training, i.e., participants were exposed to both conforming and non-conforming instances, and were told which were which, whereas Experiment 1, like other phonotactic experiments, used unsupervised training. Second, participants in the classical experiments were instructed to search for a pattern or rule; those in Experiment 1 and other phonotactic experiments were not. Both supervised training and explicit rule-seeking are known to facilitate Type II relative to Type IV in non-linguistic learning (Love, 2002; Love and Markman, 2003; Kurtz et al., 2013). Experiment 2 therefore uses unsupervised training and no rule-seeking instructions.

Thirdly, where previous experiments used a three-dimensional stimulus space, Experiment 1 used eight dimensions, of which at most three were relevant to the pattern and the rest were irrelevant. Adding irrelevant visual dimensions can hurt performance on Type II, and even more on Type VI, relative to Type I (Kepros and Bourne, 1966), but we have no information on how irrelevant dimensions affect the relation between Type II and Types III, IV, and V in human learning (for effects in GMECCS, see Footnote 7, above). The effects of irrelevant dimensions may thus reinforce or counteract those of supervised training and explicit rule-seeking. To match this aspect of Experiment 1, Experiment 2 uses an eight-feature visual stimulus space analogous to that of Experiment 1.

Fourthly, the features in Experiment 1, unlike those in the classic experiments, were not entirely orthogonal: Some pairs of features were more closely linked than others owing to the prosodic and featural structure of pseudo-word stimuli. For instance, the height of the first vowel has a structural relationship to the height of the second vowel, or to the backness of the first vowel, which it does not have to the backness of the second vowel or the voicing of the second consonant. These factors are known to affect the relative difficulty of Types II and IV (Love and Markman, 2003). The stimuli in Experiment 2 are therefore designed to have internal structure that is analogous to segments, syllables, and autosegmental feature tiers.

Table 7: Correspondence between cake and word features.

Feature	Stimulus segment			
	σ_1		σ_2	
	C_1	V_1	C_2	V_2
voiced	±		±	
Coronal	±		±	
high		±		±
back		±		±

Feature	Nonlinguistic analogues			
	Layer 1 (Bottom)		Layer 2 (Top)	
	Candy 1	Body 1	Candy 2	Body 2
Diamond candy	±		±	
Blue candy	±		±	
White icing		±		±
Brown batter		±		±

6.1 Method

6.1.1 Participants

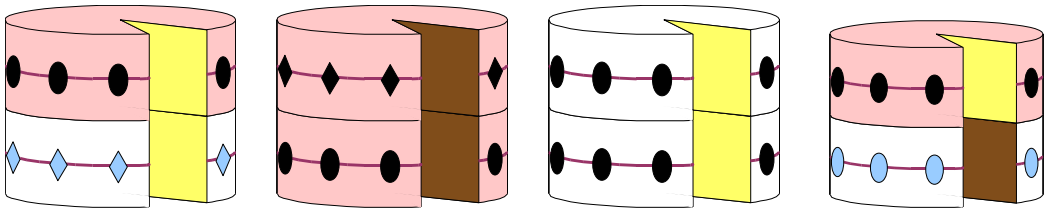
Volunteers were recruited from the same population as in Experiment 1, at the same rate of pay. Again, there were 24 participants in each Type condition, for a total of 144. Another 15 people took part, but their data could not be used (1 had non-native English, 3 failed to meet the 12-out-of-32 accuracy criterion, and 12 were lost to equipment or software failure).

6.1.2 Stimuli

The objective of stimulus design for this experiment was to create a stimulus space with the following properties, which are meant to make Experiment 2 analogous to Experiment 1. (A) Each word stimulus from Experiment 1 should have a unique visual analogue in Experiment 2. (B) The internal prosodic structure of the words should be reflected in the analogues. In particular, there should be an analogue of the segment, grouping two features together, and an analogue of the syllable, grouping two segment analogues together. (C) The featural tier structure of the words should be reflected in the analogues: Each type of feature should occur twice, once in each syllable analogue. Each type of feature should occur exclusively in the consonant analogues or in the vowel analogues. Properties (B) and (C) introduce non-orthogonality between visual features that is analogous to a non-orthogonality between phonological features in Experiment 1. (D) The objects should be relatively "natural" in the sense of being of a recognizable and familiar sort (the same way that nonsense words are a recognizable and familiar thing).

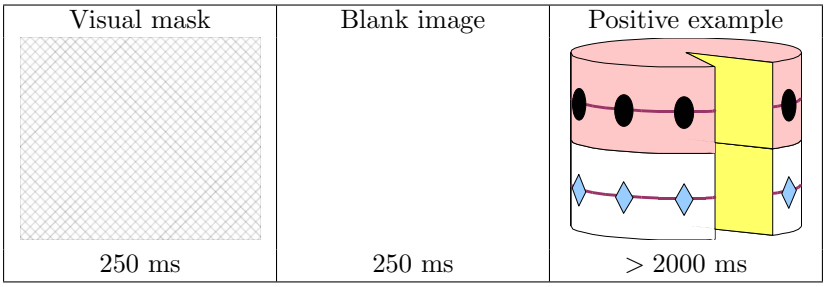
Stimuli were fancy cakes. The cakes were organized into two layers (the highest level of grouping, analogous to syllables), and within the layers, into the body of the layer and the stuck-on decorations (analogous to vowels and consonants, respectively). Table 7 shows the analogy. Some examples of words and their analogous cakes are shown in Fig. 12. As in Experiment 1, each participant was trained on a different instantiation of the SHJ pattern type appropriate for their condition.

Figure 12: Examples of corresponding cakes and words.



[dikæ] [kɔgɔ] [kiki] [tugæ]

Figure 13: Sequence of events for the familiarization phase of Experiment 2.

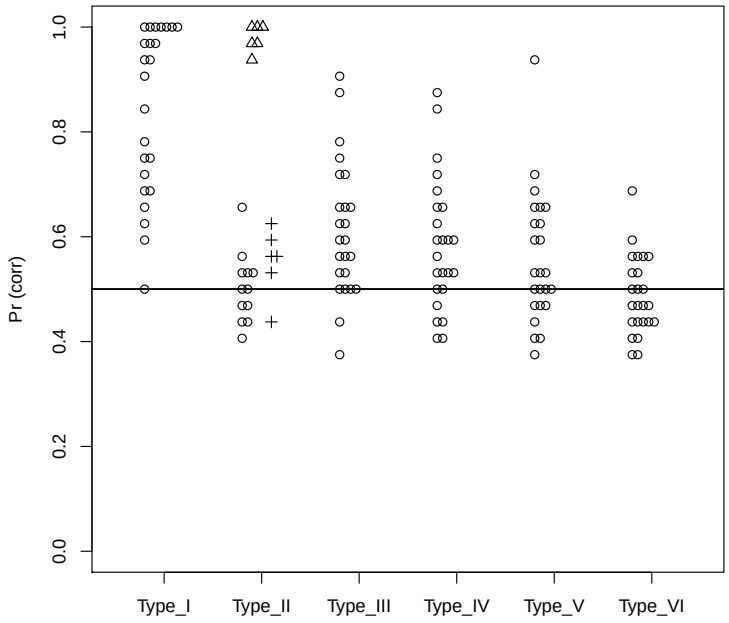


6.1.3 Procedure

The procedure was modified minimally from that of Experiment 1; only the differences will be described here. Participants were told that they would be learning to recognize "a particular style of fancy cake". They would first study cakes made in this style, then they would be "tested on how well you can recognize them." In the familiarization phase, each participant viewed 32 pattern-conforming cakes, one at a time, in random order four times. They could view each cake for as long as they liked before proceeding to the next cake. Cakes were separated by a 250-ms visual mask and 250-ms blank image.

The test phase consisted of 32 two-alternative forced-choice trials, each with one pattern-conforming cake and one non-conforming cake. All images were 7 cm wide by 5.5 cm high, and were displayed on a screen at a self-selected distance (about 45 cm) from the participant's eye. Each trial began with a 1000-ms fixation point, followed by a 250-ms blank screen. One of the cakes was then exposed for 2000 ms. A 250-ms visual mask and 250-ms blank image followed, and then the other cake was presented for 2000 ms, followed again by a 250-ms visual mask. The display region then remained blank until the participant had responded (by clicking "1" or "2" on buttons displayed below it). The next trial began as soon as the participant had responded, but not less than 250 ms after the disappearance of the visual mask at the end of the current trial.

Figure 14: Individual participant performance (proportion pattern-conforming responses) for Experiment 2. Plotting symbols for Type II: + = all relevant features are in the same segment analogue (Type II-ss); △ = all relevant features are of the same type (agreement/disagreement pattern, Type II-sg); ∘ = other.



6.2 Results and discussion

Individual participant means are shown in Table 8 and in Fig. 14. The same analysis procedure was followed as for Experiment 1. The fixed-effects parameters of the fitted model are shown in Table 9. Performance in the reference group (Type II with the two critical features belonging to different types and different layers) did not differ significantly from chance. Familiarization on Type III significantly increased the odds of a pattern-conforming response by a factor of 1.61 ($= e^{0.47634}$). Pattern-conforming responses were also more likely in the Type IV and Type V conditions, but the differences did not reach significance. As far as the relative difficulty of Type II compared to Types III, IV, and IV goes, participants in this experiment performed more like those in Experiment 1 than like those in the classic Shepard experiments: There was no evidence that Type II was easier than Types III, IV, and V, and in fact there was positive evidence that Type II was *harder* than Type III. These results are consistent with the use of a cue-based learning procedure of the sort described in Section 3. and inconsistent with that of a rule-based one.

However, the results also differed in some ways from those of Experiment 1. Seven of the 24 Type I

Type	I	II	II	II	III	IV	V	VI
Plotting symbol	∘	∘	△	+	∘	∘	∘	∘
Mean	0.845	0.503	0.979	0.552	0.613	0.585	0.557	0.489
SD	0.160	0.067	0.026	0.065	0.130	0.126	0.125	0.075
N	24	12	6	6	24	24	24	24
s.e.m.	0.032	0.019	0.011	0.027	0.027	0.026	0.026	0.015

Table 8: Mean proportion correct in each Type condition of Experiment 2, with standard deviations, cell sizes, and standard errors. Plotting symbols for Type II: + = all relevant features are in the same segment analogue; △ = all relevant features are of the same type (agreement/disagreement pattern); ∘ = other.

participants chose the pattern-conforming response on all 32 of the test trials, and fully half of the participants chose it on at least 30 of them. In fact, Type I performance was very significantly better than that in all other Type conditions as well, as shown by the fact that the Type I model coefficient was four standard errors above the next-highest coefficient (that of Type III). Near-perfect Type I performance, beside poorer performance on higher types, is characteristic of rule induction rather than cue-based learning (Shepard et al., 1961; Smith et al., 2004). The theoretical reason is that Type I patterns are solved by single-feature rules, which, being less syntactically complex than any other kind of rule,[12] are found more quickly, and, once found, enable perfect performance. It has been found empirically that task conditions which favor rule learning, such as explicit instructions to seek a rule, lead to a bimodal distribution of participant performance, with "solvers" near 100% accuracy and "nonsolvers" near 50% (Kurtz et al., 2013).

An alternative cue-based account might be possible, but would have to be specially engineered for the task. One approach to privileging Type I over the others would be to inhibit the two- and three-feature constraints in GMECCS or another model. However, some explanation would then be required for the empirical finding, in Figure 15, that the preference between the two central stimuli in Type IV — which is supported by six partially-valid one-feature (face) constraints — is so much weaker than the preference in Type I, which is supported by two fully-valid one-feature constraints, and why it is little better than Type II, which is supported by no one-feature constraints at all.

Finally, there were differences related to the internal structure of the stimuli. In Experiment 1, performance on Type II patterns improved significantly when both relevant features were in the same consonant or same vowel, but Type II performance was not significantly affected when both relevant features were of the same type. The reverse was true for Experiment 2: Type II performance was not affected by whether the two relevant features were in the same consonant or vowel analogue, but improved significantly when both

[12]Since the definition of syntactic complexity is a model parameter, it would in principle be possible to construct a rule-based model which tried two- or three-feature hypotheses first. However, such a model would incorrectly predict that Type II would be easier than Type I, whereas the task is to explain why Type I is near-perfect and easier than Types II–VI. In other words, we are not saying that *every* rule-based model predicts near-perfect Type I performance beside mediocre performance on higher types; we are saying that the rule-based framework offers a natural explanation, without special pleading.

belonged to the same feature type (e.g., "same icing on both layers", "different batter above and below", etc.).

Coefficient	Estimate	SE	z	$\Pr(>\|z\|)$
(Intercept)	−0.17464	0.17077	−1.023	0.3064
I	1.85092	0.21958	8.429	< 0.0001
III	0.48634	0.20646	2.356	0.0185
IV	0.35657	0.20607	1.730	0.0836
V	0.24300	0.20586	1.180	0.2378
VI	−0.05394	0.20544	−0.263	0.7929
Both Same Genus	3.94839	0.58225	6.781	< 0.0001
Both Same Segment	0.21881	0.29093	0.752	0.4520
Redup	−0.23880	0.09933	−2.404	0.0162
CorrFirst	0.35885	0.06437	5.574	< 0.0001

Table 9: Summary of the fixed effects in the mixed logit model for Experiment 2 ($N = 4586$ observations, log-likelihood $= -2798$).

Both of these differences (the same-genus effect and the very good performance on Type I) may be due to facilitation of rule extraction by the higher verbalizability of the cake features compared to the phonological features. That is, subjects could have used rules to solve those types in the experiment that afforded compact verbal predicates, like "top layer is chocolate" (Type I) or "different icing on the two layers" (same-genus Type II) (Ciborowski and Cole, 1973; King and Holt, 1970; Lewandowsky et al., 2012; Kurtz et al., 2013). Other pattern types, which did not afford verbally simple rules, may have been solved using a cue-based procedure.[13]

Previous research found that naïve participants in inductive category learning experiments can verbalize rule hypotheses explicitly, either trial by trial (Bruner et al., 1956; Conant and Trabasso, 1964; Peters and Denny, 1971) or in post-experiment debriefing (Shepard et al., 1961; Gottwald, 1971; Smith et al., 1993). For example, participants trained inductively on two different logical rule schemas (e.g., X AND Y vs. X OR Y) can identify which schema correctly classifies a new set of stimuli with different features (Haygood and Bourne, 1965; King, 1966), and participants trained inductively on a rule-described pattern show positive transfer to inductive learning on a new stimulus set where the same rule is instantiated by different features (Shepard et al., 1961; Haygood and Bourne, 1965; Bourne and O'Banion, 1971). The verbal complexity of participants' rules can be correlated with the difficulty of the classification problem (Shepard et al., 1961; Ciborowski and Cole, 1973) (although the solution rate can exceed the correct verbalization rate, Ciborowski and Cole 1971). It is therefore reasonable to expect easily-verbalizable stimuli to facilitate rule extraction

[13]The coexistence of two functionally (and even neurologically) distinct learning systems in the same domain is not by any means a novel proposal; see, for example, Ashby et al. (1998) and Maddox and Ashby (2004) in vision, and Ullman (2004) in language and Wong et al. (2013) in artificial phonology learning in particular; for a critical review of multiple systems models of category learning, see Newell et al. (2011).

(Kurtz et al., 2013).

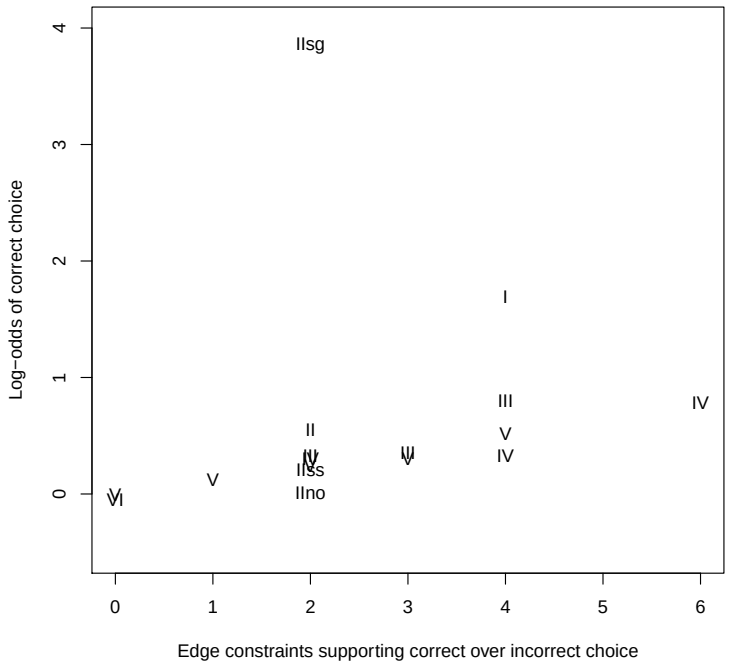
In the phonological experiment, explicit rule-based learning was not available for any of the patterns because phonological features are hard for untrained participants to verbalize. Additionally, extraction of rules (whether explicit or implicit) may also be inhibited by perceptual integrality between stimulus dimensions, which prevents selective attention to individual features. Evidence for this comes from a study by Nosofsky and Palmeri (1996) who replicated the SHJ experiment using features defined on integral dimensions (hue, brightness, and saturation). They found that Type II was now the second-most-difficult condition after Type VI. Note that although there is some evidence for integrality in the phonological features, at least in the sense that irrelevant variation in one segmental position can delay classification responses to another position (Kapatsinski, 2011), it is unlikely that this is responsible for the difficulty of Type II in Experiment 1. If that were the source of the Type II difficulty, removing the integrality ought to improve relative performance on Type II. However, although the cake features are almost paradigmatically separable, being distinct in shape, color, and location, Type II is as difficult as ever.

To sum up this discussion, had Experiment 2 yielded Shepard-like results, the difference from the results of Experiment 1 would have strengthened the hypothesis that phonotactic patterns are learned using different cognitive processes from analogous visual patterns. The actual results are in fact consistent with a cue-based learning process being available in both domains. This is confirmed when we examine the effect of edge constraints on two-alternative forced-choice responses, shown in Fig. 15. Except for Type I and the same-feature-type subtype of Type II, performance depends on how many edge constraints favor the pattern-conforming stimulus (compare Fig. 11). The results of Experiments 1 and 2 together therefore support the hypothesis that both a rule-based system and a cue-based system might be used by the learner depending on the nature of the stimuli and task conditions (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al., 2012), and that at least cue-based learning can be applied to both phonotactic and visual patterns.

7 Experiment 3: Learning with feedback

Experiment 1 showed that the behavior of participants in a standard phonotactic-learning experiment is well modelled by the cue-based learner IMECCS. Experiment 2 showed that a visual analogue evokes cue-based learning in some conditions, but rule-based learning in others, thereby confirming that the cue-based performance in Experiment 1 depends at least in part on the content of the stimuli, and is not an inevitable consequence of the experimental paradigm (unsupervised familiarize-and-test learning with multiple irrelevant features). Two main possibilities remain. One is that all phonotactic learning takes place via cue-based

Figure 15: Test performance (log-odds of probability of a pattern-conforming response) as a function of the number of edge constraints supporting the pattern-conforming stimulus over the non-conforming one.



processes, i.e., there really is only a single learning process for acquiring phonotactic patterns. Alternatively, it may be that rule-based learning is also possible for phonology, given the right circumstances (which Experiment 1 did not provide). Experiment 3 is designed to address these possibilities.

One factor that may facilitate rule-based learning relative to cue-based learning is supervised training (Love, 2002; Maddox and Ashby, 2004), i.e., a regime in which the participant explicitly sorts stimuli into categories and receives right/wrong feedback. This paradigm has also been used in studies of phonological learning, though it is less common than unsupervised training (Schane et al., 1974; LaRiviere et al., 1974, 1977; Coberly and Healy, 1984; Pycha et al., 2003, 2007). Experiment 3 is designed along similar lines. Participants are exposed to both "words" and "non-words" of the language, and learn by trial and error which are which. They are then tested using the same kind of test phase as in Experiment 1. If supervised training facilitates rule-based learning relative to cue-based learning, Experiment 3 should show improvements, compared to Experiment 1, in Types II and I relative to Type IV.

7.1 Method

Participants in Experiment 1 received 128 familiarization trials in which 32 different pattern-conforming words were presented four times each. In Experiment 3, participants likewise received 128 familiarization trials in which 32 different words were presented four times each, but in this experiment, 16 of the words were pattern-conforming and 16 were non-conforming.

7.1.1 Participants

To focus on the II *vs.* IV comparison, Types III, V, and VI were omitted entirely. There were 6 participants in each sub-type of Types I, II, and IV defined by having (vs. lacking) two critical features in the same segment and having (vs. lacking) two critical features of the same genus (for a total of 32 participants). (As in Experiments 1 and 2, preliminary analysis showed that performance on the sub-types differed only within Type II, so they were collapsed together for Type IV.) Participants were 48 paid volunteers recruited from the UNC university community. They were self-screened for normal hearing and native English. Data from 3 more participants was replaced because of equipment failure.

7.1.2 Stimuli

The stimuli were identical to those used in Experiment 1. As in Experiments 1 and 2, each participant received a different, randomly-generated instantiation of the SHJ pattern that was appropriate to their condition.

7.1.3 Procedure

The equipment and testing environment were as in Experiments 1 and 2. As in Experiment 1, participants were told that they would be learning to recognize words in an artificial language, and would be tested later on how well they could recognize them. On each trial of the study phase, the participant heard a stimulus word and repeated it aloud, then mouse-clicked one of two on-screen buttons marked "Yes" or "No". Feedback for a correct answer was a bell sound; for an incorrect one, a buzzer. Instructions and procedure for the test phase were identical to those used in Experiment 1. No feedback was given during the test phase.

7.2 Results and discussion

7.2.1 Test phase

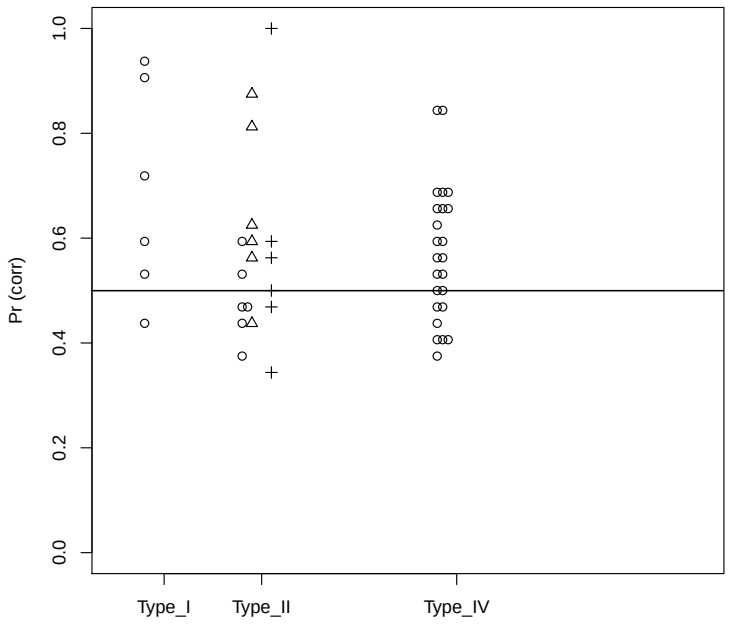
Results from the test phase are shown in Table 10 and Fig. 16. One symptom of rule-based learning — the perfect or near-perfect Type I performance that occurred so often in Experiment 2 — is conspicuously absent here. The mixed-effects logistic regression model for the results is shown in Table 9. Performance in the basic Type II condition did not differ significantly from chance, nor from performance in the Type IV condition. Type I performance was significantly above the Type II baseline. The only other significant effect of a critical variable was that of *Both Same Genus*, indicating that those Type II patterns which were based on agreement or disagreement between two instances of the same feature (e.g., the voicing features in the two consonants) elicited more pattern-conforming test-phase responses.

Type	I	II	II	II	IV
Plotting symbol	○	○	△	+	○
Mean	0.688	0.479	0.651	0.578	0.570
SD	0.203	0.077	0.163	0.224	0.129
N	6	6	6	6	24
s.e.m.	0.083	0.031	0.067	0.091	0.026

Table 10: Mean proportion correct in each Type condition of Experiment 3, with standard deviations, cell sizes, and standard errors. Plotting symbols refer to Fig. 16.

If replacing unsupervised with supervised training facilitates rule-based learning of phonotactic patterns, then substituting supervised for unsupervised training ought to improve performance on Type II patterns relative to Type IV. This prediction was tested statistically by fitting another mixed-effects logistic regression model to the data from Experiment 3 and from the Type I, II, and IV conditions of Experiment 1, with Experiment (unsupervised vs. supervised) as a factor (Table 12). The large and significantly negative main effect of *Supervised* indicates that performance on the basic Type II pattern was *lower*, not higher, when

Figure 16: Individual participant performance (proportion pattern-conforming responses) for Experiment 3. Plotting symbols: + = all relevant features are in the same segment analogue; △ = all relevant features are of the same genus (agreement/disagreement pattern); ∘ = other.



Coefficient	Estimate	SE	z	$\Pr(>\|z\|)$
(Intercept)	−0.1772	0.2780	−0.638	0.523791
I	0.9968	0.3938	2.531	0.011377
IV	0.3955	0.3053	1.295	0.195171
Both Same Genus	0.7853	0.3901	2.013	0.044085
Both Same Segment	0.4978	0.3896	1.278	0.201373
Redup	−0.6272	0.1624	−3.863	0.000112
CorrFirst	0.1726	0.1072	1.610	0.107506

Table 11: Summary of the fixed effects in the mixed logit model for Experiment 3 ($N = 1536$ observations, log-likelihood = −1002). Type II is the reference category.

supervised training was used, whereas Type IV performance was not significantly reduced. The only other statistically significant difference between the supervised and unsupervised experiments was that Type II agreement and disagreement patterns elicited significantly better performance after supervised than after unsupervised training, as seen in the significant positive effect of *Supervised* × *Both Same Genus*.

Coefficient	Estimate	SE	z	$\Pr(>\|z\|)$
(Intercept)	0.329498	0.059838	5.506	< 0.0001
I	0.681943	0.116145	5.871	< 0.0001
IV	0.516977	0.114390	4.519	< 0.0001
Both Same Segment	0.520218	0.215154	2.418	0.01561
Both Same Genus	−0.073559	0.210144	−0.350	0.72631
Supervised	−0.524418	0.182644	−2.871	0.00409
CorrFirst	0.244593	0.054265	4.507	< 0.0001
Redup	−0.721768	0.085395	-8.452	< 0.0001
Supervised × *I*	0.315515	0.283824	1.112	0.26629
Supervised × *IV*	−0.154627	0.230952	−0.670	0.50316
Supervised × *Both Same Segment*	0.004564	0.332755	0.014	0.98906
Supervised × *Both Same Genus*	0.768695	0.332198	2.314	0.02067

Table 12: Summary of the fixed effects in the mixed logit model comparing Experiment 3 with Experiment 1 ($N = 6144$ observations, log-likelihood $= -3912$). The reference category is the unsupervised basic Type II.

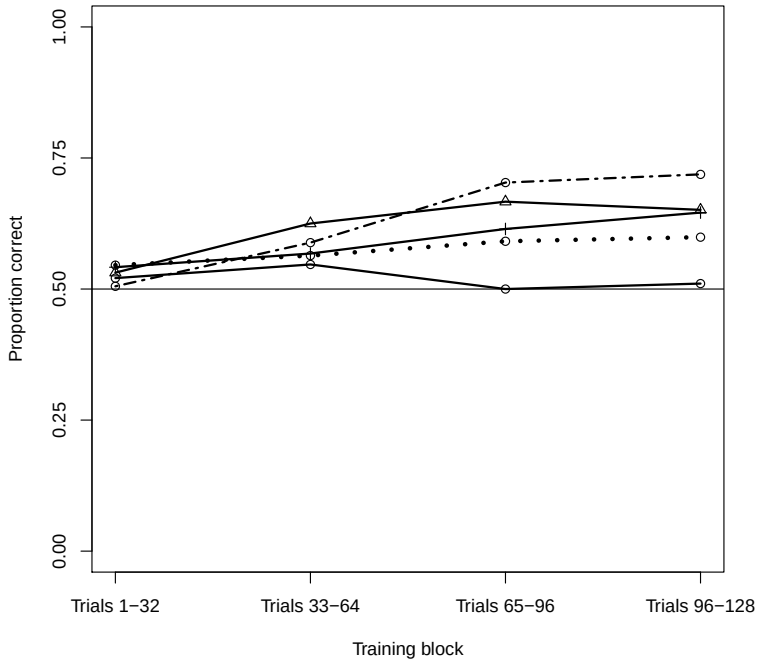
7.2.2 Training phase

The test-phase results seem to contradict the hypothesis that supervised training facilitates rule-based over cue-based learning of phonotactic patterns, since Type II performance became significantly worse rather than better after supervised training. However, it is still possible that a Type II advantage was present in the training phase, but did not carry over into the test phase (which required generalization outside of the training set.)

To test this possibility, the 128 training trials were divided into four 32-trial blocks, and average performance (proportion correct) was calculated for each block, as shown in Fig. 17. Participants in the Type I condition learned faster and reached a higher level of performance than the participants in the Type II and Type IV conditions. A logistic-regression model was fitted to the data (Fig. 13). The fixed effects were the same as in the analyses of Experiments 1 and 2 except for the addition of a Block variable (with Block 0 being Trials 1–32) and interactions with it. The `CorrFirst` variable was also dropped, since it is only meaningful for two-alternative forced-choice trials. There was again a random intercept for each participant.

Fig. 17 shows that essentially no learning took place in the basic Type II condition, an observation which is borne out by the smallness and nonsignificance of `Block` in Table 13. Learning in the Type IV condition was not significantly faster than that (the `Block` × `IV` term was positive, but not significantly so). The

Figure 17: Aggregate learning curves (mean proportion correct responses) for the training phase of Experiment 3. Line types: Dash-dotted = Type I; solid = Type II; dotted = Type IV. Plotting symbols: + = all relevant features are in the same segment; △ = all relevant features are of the same genus (agreement/disagreement pattern); ∘ = other.



significant *Block* × *I* interaction shows that learning was faster in the Type I condition than in the basic Type II condition. Learning was significantly faster for Type II patterns when the two features belonged to the same segment or same feature genus, as shown by the significant interactions of *Both Same Genus* and *Both Same Segment* with *Block*. In the training phase, unlike in the test phase, participants showed no significant preference for, or aversion to, "reduplicated" stimulus words like [gaga].

Thus, neither the training-phase nor the test-phase results provide any evidence that supervised training facilitates the learning of Type II phonotactic patterns relative to Type IV patterns. We note that this is consistent with the predictions of GMECCS. GMECCS can be trained in a supervised mode by simply adding the category label (positive or negative) as a ninth feature. The classification task can be modelled as a choice between correctly- and incorrectly-labelled versions of the same eight-bit stimulus.[14] As Fig. 18

[14]This has the effect of treating a classification task as a stimulus-completion or inference task. GMECCS at present does not distinguish classification from inference, although some other models do (Love et al., 2004). There is evidence that human learners in fact treat the two differently, and that, in particular, switching from classification to completion can boost performance on linearly-separable patterns relative to non-linearly-separable ones (Yamauchi and Markman, 1998; Yamauchi et al., 2002; Markman and Ross, 2003). Hence, an inference version of Experiment 3 would presumbaly find an even larger advantage for Type IV over Type II. Thus, even if the nine-feature GMECCS is actually a more appropriate model of inference

Coefficient	Estimate	SE	z	$\Pr(>\|z\|)$
(Intercept)	0.11630	0.19808	0.587	0.5571
I	−0.07080	0.28213	−0.251	0.8019
IV	0.06393	0.22147	0.289	0.7728
Both Same Segment	0.03580	0.28072	0.128	0.8985
Both Same Genus	0.15148	0.28258	0.536	0.5919
Block	−0.03165	0.06476	−0.489	0.6251
Redup	0.15143	0.10839	1.397	0.1624
I × *Block*	0.38321	0.09605	3.990	< 0.0001
IV × *Block*	0.11001	0.07265	1.514	0.1299
Both Same Segment ×*Block*	0.18510	0.09301	1.990	0.0466
Both Same Genus × *Block*	0.21351	0.09459	2.257	0.0240

Table 13: Summary of fixed effects for the mixed-logit model for the training phase of Experiment 3 (6148 responses from 48 participants; log-likelihood = −4087). The reference category is the first 32-trial block in the basic Type II condition.

shows, the predicted difficulty orders remain the same as in the unsupervised case (Fig. 8).

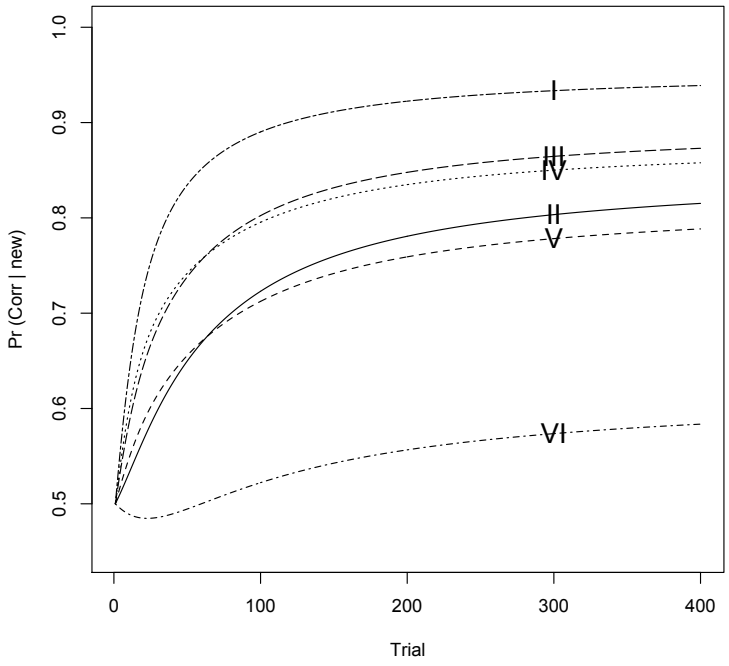
8 General Discussion

Cue-based learning models have established and expanded their foothold among linguistic theories of phonological pattern learning, while at the same time they have fallen out of favor among psychological theories of general pattern learning. This divergence is driven in part by differences in the kind of problems that are studied in the visual and phonological domains. Studies of human visual pattern learning have tended to focus on structural effects under supervised training of patterns in low-dimensional spaces with easily-verbalizable feature dimensions. In phonology, much of the interest has been in comparing different featural instantiations of structurally isomorphic patterns, using unsupervised training in higher-dimensional spaces with features that are hard for naïve participants to verbalize. The visual studies have also tended to focus on classification performance on the training stimuli, whereas the phonological studies have tended to focus on generalization to stimuli outside the training set.

The present study, we think, points towards a rapprochement between the two lines of research, focusing on both the commonalities and the differences in pattern learning across domains. Experiments 1 and 2 used the SHJ family of pattern structures from the visual literature, instantiated as analogous phonological and visual patterns, and presented to participants using the unsupervised generalization paradigm common in phonology. Pattern-structure effects in the two domains resembled each other and differed from the classic SHJ difficulty order, thereby corroborating the conclusions of Kurtz et al. (2013) about the fragility of that order (neither supervised training nor verbalizable stimuli is sufficient to elicit it), and extending them to

than of classification, it is still supported (albeit more weakly) by the results of Experiment 3.

Figure 18: Predicted generalization performance (probability of correctly classifying a novel stimulus) for 8-feature GMECCS under supervised training (average of 250 replications per Type; learning rate $\eta = 0.01$). Standard deviation for each curve at every time point shown is less than 0.015 for Types I–V, and less than 0.022 for Type VI. Standard error of the mean is standard deviation divided by $\sqrt{250}$. Chance performance is 0.5.



all six SHJ pattern types. Participant behavior in both domains was well described by GMECCS, a cue-based learner which fuses the (visual) Configural Cue Model (Gluck and Bower, 1988a) with a Maximum Entropy phonotactic learning framework (Goldwater and Johnson, 2003; Jäger, 2007; Hayes and Wilson, 2008). Although Experiment 2 found evidence consistent with rule-based learning (in the sense of a preference for hypotheses involving fewer features) in some of the simpler visual patterns, no such evidence was found for phonotactic patterns, even when supervised training was used (Experiment 3).[15]

Prototype models, which represent a category as an average across its exemplars, are also challenged by these results. Since the decision bound in a prototype model is a hyperplane perpendicular to the line that joins the two opposing prototypes, prototype models have difficulty with categories that are not linearly separable (Medin and Schwanenflugel, 1981). The non-linearly-separable Type II is thus correctly predicted to be more difficult than the linearly-separable Type IV, but Types III and V, which are not linearly separable, are incorrectly predicted to have difficulty similar to that of Type II, rather than that of Type IV.

Another major class of categorization models eliminates abstraction over training instances and instead derives categorization behavior from similarity to stored exemplars. (For a recent review of exemplar models in psychology, see Kruschke 2008; in phonology, Kirchner et al. 2010). Abstraction in GMECCS resides in the constraints, whose weights represent the model's confidence in the generalizations they express. Since the constraint weights can be eliminated from the model (Appendix A), the only necessary state variables in GMECCS are the estimated probabilities of the individual stimuli. That raises the question of whether GMECCS, and other Max Ent models described by Equations 1–5 (i.e., Replicator learners), is an exemplar model in disguise. This possibility is reinforced by the observation that the predictions of GMECCS are largely determined by the neighborhoods of the stimuli. In Section 4.1 we presented this fact in terms of stimuli sharing the support of edge or face constraints, but it can equally well be viewed as a neighborhood effect, with the constraints playing the role of a similarity function.[16]

[15]Another property which has been used to distinguish rule-based from (some) cue-based learning models is the use of "algebraic" variables (Marcus et al., 1999; Berent et al., 2012). The idea is that an appropriate rule-based learner, trained on the sequences AA, BB, CC, can form the generalization xx, i.e., "same letter twice", and can recognize DD as conforming to it, whereas connectionist or exemplar models could not. Participants in all three of our experiments were sensitive to syllable reduplication (the *Redup* factor was always significant). Those in Experiments 2 and 3 were additionally sensitive to agreement/disagreement patterns (the *Both Same Genus* factor), and those in Experiment 1 showed a non-significant trend in that direction. In the reduplication case, it is clear that participant responses are affected by abstract relationship within the stimulus, since any reduplicated test word was not heard in training. In the agreement/disagreement case, the feature sequences that were preferred in testing (e.g., [+voice]…[+voice] and [–voice]…[–voice]) were actually experienced in training, and need not necessarily involve mental variables. Even if they do, those results are neutral between a rule- and cue-based interpretation. If RULEX, for instance, is augmented to allow variables in its two-feature rules, GMECCS can be augmented to allow variables in its two-feature constraints (Moreton, 2012). If reduplication is allowed to be a dimension of similarity, even prototype and exemplar models may be able to account for sensitivity to it.

[16]In the Replicator update rule derived in the Appendix (Equation 16), the (i, j)-th entry of the matrix $C^T C$ is the dot product of the score vectors of the stimuli x_i and x_j, i.e., $\sum_{k=1}^{n} c_k(x_i)c_k(x_j)$. In the special case of GMECCS, the constraints are all binary, so the (i, j)-th entry of $C^T C$ is simply the number of constraints which have the value 1 for both x_i and x_j. The matrix $C^T C$ can therefore be viewed as a similarity matrix defined by the constraints. For the GMECCS constraints, the number of constraints which give 1 to both x_i and x_j grows as two to the power of the number of shared features, so C

However, there are also major differences between the Replicator form of GMECCS and exemplar models. Most notably, the GMECCS stimulus-probability estimates do not constitute a record of the model's experience. It is not even possible to tell from them whether a trained GMECCS has actually encountered a given stimulus, since there is a "node" (i.e., a characteristic constraint) for every *possible* stimulus, experienced or not. Nor does GMECCS generalize by comparing a new stimulus to similar old ones; rather, it compares the probability estimates for the two different labellings of the new stimulus, and ignores all other stimuli.

It is a separate question whether the learning curves of GMECCS, or of the humans in these experiments, can be replicated by an exemplar model with an appropriate similarity function.[17] When attentional learning is turned off, the exemplar model ALCOVE predicts SHJ Types III and IV to be easier than Types II and V in supervised training without generalization (Kruschke, 1992, Fig. 5, p. 28). However, that has the side effect of collapsing Type III with Type IV, and Type II with Type V. Restoring a moderate amount of attentional learning makes Type II easier than Type V. These are said (p. 27) to be the only two orderings that can be obtained. However, we do not know what ALCOVE (or other exemplar-based models) predicts for generalization from unsupervised training.

In this paper we have explored the predictions of Gradual MaxEnt learning with very simple constraints. Much of the appeal of MaxEnt models for phonology is that they can be used with constraints of arbitrary complexity, and thus provide a probabilistic approximation of any analysis in the popular Optimality Theory framework. Any pattern over a finite set of data analyzed in standard Optimality Theory with a ranking of a set of constraints can also be modeled with a weighting of the same constraints in Harmonic Grammar (Prince and Smolensky 1993, 236; Smolensky and Legendre 2006; Pater 2009, 1007). That pattern can therefore be generated with arbitrarily high probability using the probabilistic MaxEnt variant of Harmonic Grammar proposed by Goldwater and Johnson (2003). MaxEnt models can also be applied to derivational models of grammar with an unbounded number of mappings between representational levels, which yield patterns that go beyond those expressible with standard Optimality Theory's two-level mappings (see Staubs and Pater 2014 on MaxEnt learning of Harmonic Serialism, and Johnson 2013 on MaxEnt learning of Minimalist syntactic grammars). This all means that Gradual MaxEnt learning can be applied to generate predictions for laboratory learning of patterns of complexity similar to the range found in natural languages.

Learning difficulty depends on the constraint set as well as the update algorithm. There are two main proposals as to the source of the constraint set. One is that it is pre-specified and does not change during learning. An example in the MaxEnt framework would be Goldwater and Johnson (2003)'s phonological

measures featural similarity between stimuli.

[17] Because the similarity matrix C^TC in Equation 16 is C times its own transpose, it is symmetric; Stimulus x_i is as similar to Stimulus x_j as x_j is to x_i. Evidence that concept learning is described by an asymmetric similarity matrix would therefore tell against GMECCS, with the unbiased constraint set or with any other constraint set.

learner, which is equipped with specifically phonological constraints such as `*Lapse` ("No consecutive unstressed syllables"). Alternatively, the constraints may be induced from the data, such that the constraint set changes while the weights are being learned (Della Pietra et al., 1997). The phonotactic MaxEnt learner of Hayes and Wilson (2008) follows this procedure, generating constraints according to a schema and preferentially adding ones that reduce error. GMECCS takes an intermediate approach by pre-loading the constraint set with all constraints that conform to the conjunctive schema.[18] This choice makes GMECCS simple, general, and analytically tractable, but it is almost certainly wrong empirically, and there is evidence that it is wrong enough to matter. For one thing, the order of difficulty of two (non-SHJ) visual concepts may vary depending on whether participants know beforehand which features are relevant (Giambra, 1970). For another, categorization may make use of configural cues which are learned in the lab or in nature (Pevtzow and Goldstone, 1994; Ross, 1996; Markman and Ross, 2003).[19] Although our experiments used features which are presumably familiar to participants (phonetic features of their native language, and simple visual or culinary features), participants may have had pre-existing representations for some configurations of those features (e.g., the conjunction of height and backness that makes the phoneme [u], or the conjunction "blue diamond") but have had to induce others (e.g., the conjunction of the voicing of the initial consonant with the height of the final vowel, or between diamond-shaped candy and chocolate batter), thereby disadvantaging some patterns relative to others. For example, the unexpectedly poor human performance on basic Type II patterns in all three experiments, and the better performance on same-segment (Experiment 1) or same-genus (Experiment 2) Type II, might be due to a lack of pre-existing predicates for generic Type II relations (Moreton, 2012).

Natural-language phonological patterns are cultural products whose persistence depends in part on the ability of each generation to learn them from their elders. In this, they differ from many other kinds of learned pattern which are stabilized by fixed properties of the environment (e.g., the constellation of features separating sheep from goats). When other factors are controlled, phonological patterns which are easier to learn ought to be more frequent across historically-distinct languages (Bell, 1970, 1971; Greenberg, 1978).

This hypothesis has already been investigated in the search for "substantive" inductive biases, i.e., differences in learnability when the same structural pattern is instantiated by different phonetic features. The search is motivated by the observation that such instantiations often differ greatly in frequency across natural

[18] GMECCS inherits this approach from its connectionist predecessor, the Configural Cue Model. The network analogue of inducing constraints from the data is inducing the weights between single-feature input units and a layer of hidden units, with the network's final decision being determined by a weighting of the outputs of the hidden units. The Configural Cue Model in effect provides a pre-specified hidden unit for each combination of input features, with pre-specified and inalterable weights between each feature input unit and each hidden unit. (E.g., the hidden unit for $+++$ would have weights of 1 from each of the three input units representing $+$ in each position, and weights of 0 from the other three input units.) This eliminates the need to learn the first layer of weights (Gluck and Bower, 1988a, 187–188).

[19] The authors are indebted to [name omitted] for pointing out the relevance of this literature.

languages; e.g., the example from Section 2.1 that many languages require syllable- or word-final obstruent consonants to lack voicing, but few or none require them to possess it. The intensive search for learnability differences answering to these typological asymmetries has, when examined closely, so far found only weak and inconsistent corroboration (along with the examples cited in Section 2.1, see the review in Moreton and Pater 2012b).

When instead the features are kept constant and the pattern structure is manipulated, the effects on learnability are strong and consistent — not only with each other (as reviewed in Moreton and Pater 2012a), but, as we have now seen, also with analogous effects in non-linguistic learning. Other structural effects seen in non-linguistic domains are visible in phonology as well. An immense survey by Mielke (2004, 2008) has found that cross-linguistically, phonological patterns tend to be based on sound classes that are expressible as conjunctions, or low-order disjunctions, of phonetic features. This is consistent with a large body of non-linguistic research comparing conjunctions with other logical connectives such as disjunctions and biconditionals (e.g. Bruner et al., 1956; Neisser and Weene, 1962; Hunt and Kreuter, 1962; Conant and Trabasso, 1964; Haygood and Bourne, 1965), and there is evidence that it holds for morphological learning as well (Pertsova, 2012). Non-linguistic and phonological learning likewise share a special sensitivity to intra-dimensional Type II patterns (i.e., those based on agreement or disagreement between two features of the same genus, like vowel height harmony) compared to inter-dimensional ones (Moreton, 2008; Lin, 2009; Moreton, 2012).

The depth and pervasiveness of pattern-structure effects suggests that when other factors are controlled, pattern structure should influence natural-language typological frequency, since more difficult structures will tend to be changed or altered in transmission (Bach and Harms, 1972). Controlling these other factors is not trivial, because the innovation and extinction of phonological patterns may be skewed by articulatory and perceptual biases in the phonetic channel between speakers and hearers (e.g., Hyman, 1976; Ohala, 1993; Barnes, 2002; Blevins, 2004). A full account of typology will require modelling of not only the inductive biases and the channel biases, but of their interaction during iterated learning (Griffiths and Kalish 2007; Griffiths et al. 2008; Rafferty et al. 2012; see also Pater and Moreton 2012 for preliminary work on iterated learning with MaxEnt grammars, and for references to related work on agent-based modeling).

A Gradient-descent Max Ent and the Replicator Equation

This appendix shows that any Maximum Entropy learner that uses (unregularized) gradient descent on negative log-likelihood is equivalent, in terms of learning curves, to a weightless model in which the updates are done directly on the stimulus probabilities using a special case of the Replicator Equation from evolutionary biology. The equivalence is not limited to GMECCS; it holds if any other constraint set is substituted for the GMECCS constraints.

We will use the following notation: there are m constraints ("features" in machine-learning terms) $\{c_i\}_1^m$, m weights $\{w_i\}_1^m$, and n stimuli $\{x_j\}_1^n$. The empirical expectation of a random variable X is denoted $E_{emp}[X]$, while the learner's expectation when the weight vector $\mathbf{w} = (w_1, \ldots, w_m)$ is $E_{\mathbf{w}}[X]$. The first step in the derivation is to determine how changing a weight w_k affects the model's expectation of c_i. That expectation is just the probability-weighted sum of $c_i(x_j)$ for each stimulus x_i:

$$\begin{aligned}\frac{\partial}{\partial w_k} E_{\mathbf{w}}[c_i] =& \frac{\partial}{\partial w_k} \sum_{j=1}^{n} \Pr(x_j \mid \mathbf{w}) \cdot c_i(x_j) \\ =& \sum_{j=1}^{n} \frac{\partial}{\partial w_k} \Pr(x_j \mid \mathbf{w}) \cdot c_i(x_j) \\ =& \sum_{j=1}^{n} \left(c_i(x_j) \cdot \frac{\partial}{\partial w_k} \Pr(x_j \mid \mathbf{w}) \right)\end{aligned} \tag{8}$$

It is now necessary to determine how changing weight w_i affects the model's assignment of probability to stimulus x_j. We start with the definition of the stimulus probability in Equation 3 and differentiate it:

$$\frac{\partial}{\partial w_i} \Pr(x_j \mid \mathbf{w}) = \frac{\partial}{\partial w_i} \frac{\exp h_{\mathbf{w}}(x_j)}{Z_{\mathbf{w}}} \tag{9}$$

After applying the quotient rule, we make the substitution $\partial Z_{\mathbf{w}} / \partial w_i = Z_{\mathbf{w}} \cdot E_{\mathbf{w}}[c_i]$ (which follows straightforwardly from differentiating Equation 2), and then do a little more algebra to get:

$$\begin{aligned}\frac{\partial}{\partial w_i} \Pr(x_j \mid \mathbf{w}) =& \frac{\exp h_{\mathbf{w}}(x_j)}{Z_{\mathbf{w}}} (c_i(x_j) - E_{\mathbf{w}}[c_i]) \\ =& \Pr(x_j \mid \mathbf{w}) \cdot (c_i(x_j) - E_{\mathbf{w}}[c_i])\end{aligned} \tag{10}$$

Substituting Equation 10 into Equation 8 now yields:

$$\begin{aligned}\frac{\partial}{\partial w_k} E_{\mathbf{w}}[c_i] =& \sum_{j=1}^{n} (c_i(x_j) \cdot \Pr(x_j \mid \mathbf{w}) \cdot (c_k(x_j) - E_{\mathbf{w}}[c_k])) \\ =& \sum_{j=1}^{n} (\Pr(x_j \mid \mathbf{w}) \cdot (c_i(x_j) c_k(x_j) + c_i(x_j) \cdot E_{\mathbf{w}}[c_k])) \\ =& E_{\mathbf{w}}[c_i c_k] - E_{\mathbf{w}}[c_i] E_{\mathbf{w}}[c_k] \\ =& \mathrm{Cov}_{\mathbf{w}}[c_i, c_k]\end{aligned} \tag{11}$$

The gradient descent update rule in Equation 5 tells us how the weights change, and Equation 11 tells

us how each weight change changes the expectations. Putting these together, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[c_i] =& \sum_{k=1}^{n}\frac{\partial}{\partial w_k}E_{\mathbf{w}}[c_i]\cdot\frac{\partial w_k}{\partial t}\\ =&\eta\sum_{k=1}^{n}\mathrm{Cov}_{\mathbf{w}}[c_i,c_k]\cdot(E_{emp}[c_k]-E_{\mathbf{w}}[c_k])\end{aligned} \tag{12}$$

Now we augment the constraint set by adding, for every stimulus x_j, a "characteristic constraint" $\hat{c}_j$ whose value is some small ϵ for x_j and is 0 for any other stimulus. The characteristic constraints have three useful properties. The first is that by choosing ϵ sufficiently small, we can reduce their effect on the estimated stimulus probabilities to any desired level, and thus cause the augmented learner to approximate the behavior of the un-augmented learner as closely as we like. The second is that the probability estimate for x_j is proportional to the expected value of $\hat{c}_j$; i.e., $E_{\mathbf{w}}[\hat{c}_j]=\epsilon\Pr(x_j\mid\mathbf{w})$ (this can be seen by setting $\epsilon=1$). The third is that $\hat{c}_j$ stands in a convenient relationship to any other constraint c_i: $c_i(x_k)\hat{c}_j(x_k)=c_i(x_j)\hat{c}_j(x_j)$ if $k=j$, and is otherwise zero. Hence, $E_{\mathbf{w}}[c_i\hat{c}_j]=c_i(x_j)\cdot E_{\mathbf{w}}[\hat{c}_j]$. The covariance between a characteristic constraint and any other constraint is therefore

$$\begin{aligned}\mathrm{Cov}_{\mathbf{w}}[c_i,\hat{c}_j] =&E_{\mathbf{w}}[c_i,\hat{c}_j]-E_{\mathbf{w}}[c_i]E_{\mathbf{w}}[\hat{c}_j]\\ =&c_i(x_j)\cdot E_{\mathbf{w}}[\hat{c}_j]-E_{\mathbf{w}}[c_i]E_{\mathbf{w}}[\hat{c}_j]\\ =&E_{\mathbf{w}}[\hat{c}_j]\cdot(c_i(x_j)-E_{\mathbf{w}}[c_i])\\ =&\epsilon\cdot\Pr(x_j\mid\mathbf{w})\cdot(c_i(x_j)-E_{\mathbf{w}}[c_i])\end{aligned} \tag{13}$$

The covariance between two characteristic constraints is therefore bounded between the negligibly small quantities $-\epsilon^2$ and $+\epsilon^2$, so we can safely ignore them. Now we can substitute Equation 13 back into Equation 12 to get:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Pr(x_j\mid\mathbf{w}) =&\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\epsilon}E_{\mathbf{w}}[\hat{c}_j]\\ =&\eta\sum_{i=1}^{m}\frac{1}{\epsilon}\mathrm{Cov}_{\mathbf{w}}[c_i,\hat{c}_j]\cdot(E_{emp}[c_i]-E_{\mathbf{w}}[c_i])\\ =&\eta\cdot\Pr(x_j\mid\mathbf{w})\cdot\left(\sum_{i=1}^{m}(c_i(x_j)-E_{\mathbf{w}}[c_i])\cdot(E_{emp}[c_i]-E_{\mathbf{w}}[c_i])\right)\end{aligned} \tag{14}$$

For brevity, let $p_j=\Pr(x_j\mid\mathbf{w})$, let $q_i=E_{\mathbf{w}}[c_i]$, and let $q_i^\star=E_{emp}[c_i]$. Then Equation 14 becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}p_j=\eta\cdot p_j\cdot\sum_{i=1}^{n}(c_i(x_j)-q_i)\cdot(q_i^\star-q_i) \tag{15}$$

which completes the derivation of Equation 6.

If the model's current estimated probabilities $\{p_j\}_1^n$ for all stimuli are known, then $\mathbf{w}$ is not needed to evaluate Equation 15, because we can calculate q_i by simply adding up the probability of each stimulus times its score on c_i. To calculate $\mathbf{q}=\{q_i\}_1^m$ from $\mathbf{p}=\{p_j\}_1^n$, let C be the matrix whose jth column is the score

vector of x_j, i.e., $C_{i,j} = c_i(x_j)$. Then $\mathbf{q} = C\mathbf{p}$. If we let $\mathbf{e} = \mathbf{p}^\star - \mathbf{p}$ be the model's error, then Equation 15 yields the following update rule:

$$\Delta p_j = \eta \cdot p_j[(S\mathbf{e})_j - \mathbf{p}^T S\mathbf{e}] \tag{16}$$

where $S = C^T C$. The (i, j)-th entry of the symmetric matrix S is the dot product of the score vectors of x_i and x_j, which is a measure of similarity between the two stimuli in constraint-score space (e.g., if the constraints are binary-valued, $S_{i,j}$ is the number of constraints that give a 1 to both stimuli).

Equation (16) is a special case of the more general update rule

$$\Delta p_j = \eta \cdot p_j[\mathbf{f}(\mathbf{p})_j - \mathbf{p}^T \mathbf{f}(\mathbf{p})] \tag{17}$$

where $\mathbf{f}$ is any real vector-valued function of $\mathbf{p}$. This equation arises in ecology, genetics, and evolutionary biology under the name of the Replicator Equation.[20] In the biological interpretation, p_j is the concentration of the j-th species in a closed ecosystem, and $\mathbf{f}$ assigns a numerical fitness score to each species depending on the current distribution of species. One class of Replicator systems, in which an individual's fitness depends on a time-varying environment, but not on other individuals' fitness, is already known to be related to Maximum Entropy (Karev, 2010). The Replicator systems studied here differ in that an individual's fitness depends on others' fitness and not on a time-varying environment.

We can thus think of the weightless equivalents of GMECCS and related Max Ent models as "Replicator learners". If the fitness function $\mathbf{f}$ is chosen as in Equation 16, the Replicator learner behaves like the corresponding Max Ent learner. However, the class of Replicator learners is strictly larger than the class of Max Ent learners (for example, if the similarity matrix S in Equation 16 is not symmetric, then it cannot be expressed as $C^T C$ for any constraint set); hence, Replicator learners provide both a new way to analyze Max Ent learners, and a new way to generalize them.

[20] The authors are greatly indebted to Paul Smolensky for identifying Equation 17 as the Replicator.

References

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

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

Ashby, F. G. and W. T. Maddox (2005). Human category learning. *Annual Review of Psychology 56*, 149–178.

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

Ashley, K. C., L. Disch, D. C. Ford, E. MacSaveny, S. Parker, C. Unseth, A. M. Williams, R. Wong, and B. Yoder (2010). How many constraints are there? a preliminary inventory of OT phonological constraints. Graduate Institute of Applied Linguistics Occasional Papers in Applied Linguistics, No. 9.

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

Barnes, J. (2002). *Positional neutralization: a phonologization approach to typological patterns*. Ph. D. thesis, University of California, Berkeley.

Bell, A. (1970). *A state-process approach to syllabicity and syllabic structure*. Ph. D. thesis, Stanford University.

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

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

Blevins, J. (2004). *Evolutionary phonology*. Cambridge: Cambridge University Press.

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

Bourne, L. E. and K. O'Banion (1971). Conceptual rule learning and chronological age. *Developmental Psychology 5*(3), 525–534.

Bruner, J. S., J. J. Goodnow, and G. A. Austin (1956). *A study of thinking*. New York: John Wiley and Sons.

Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language Learning and Development 7*(4), 263–278.

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

Christiansen, M. H. and N. Chater (2008). Language as shaped by the brain. *Behavioral and Brain Sciences 31*(5), 489–509.

Ciborowski, T. and M. Cole (1971). Cultural differences in learning conceptual rules. *International Journal of Psychology 6*(1), 25–37.

Ciborowski, T. and M. Cole (1973). A developmental and cross-cultural study of the influences of rule structure and problem composition on the learning of conceptual classifications. *Journal of Experimental Child Psychology 15*(2), 193–215.

Coberly, M. S. and A. F. Healy (1984). Accessibility of place and manner features and the place/manner dissimilation principle in a learning task. *Language and Speech 27*(4), 309–321.

Coetzee, A. and J. Pater (2011). The place of variation in phonological theory. In J. Goldsmith, J. Riggle, and A. Yu (Eds.), *The Handbook of Phonological Theory* (2 ed.)., pp. 401–431. Blackwell.

Conant, M. B. and T. Trabasso (1964). Conjunctive and disjunctive concept formation under equal-information conditions. *Journal of Experimental Psychology 67*(3), 250–255.

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

Crump, M. J. C., J. V. McDonnell, and T. M. Gureckis (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE 8*(3), e57410.

Daland, R., B. Hayes, J. White, M. Garellek, A. Davis, and I. Norrmann (2011). Explaining sonority projection effects. *Phonology 29*, 197–234.

Della Pietra, S., V. Della Pietra, and J. Lafferty (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*(4), 380–393.

Dutoit, T., V. Pagel, N. Pierret, F. Bataille, and O. van der Vreken (1996). The MBROLA Project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP) 3*, pp. 1393–1396.

Evans, N. and S. C. Levinson (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences 32*, 429–492.

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

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science 12*(6), 227–239.

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

Fific, M., D. R. Little, and R. M. Nosofsky (2011). Logical-rule models of classification response times: a synthesis of mental-architecture, random-walk, and decision-bound approaches. *Journal of Experimental*

Psychology: Human Perception and Performance 117(2), 309–348.

Finley, S. and W. Badecker (2010). Linguistic and non-linguistic influences on learning biases for vowel harmony. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Sciety*, Austin, Texas, pp. 706–711. Cognitive Science Society.

Gallistel, R. (2011). Prelinguistic thought. *Language Learning and Development 7*(4), 253–262.

Giambra, L. M. (1970). Conditional and biconditional rule difficulty with attribute identification, rule learning, and complete learning task. *Journal of Experimental Psychology 86*(2), 250–254.

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

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

Goldwater, S. J. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Erkisson, and O. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pp. 111–120.

Goodman, N., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths (2008). A rational analysis of rule-based concept learning. *Cognitive Science 32*(1), 108–154.

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

Gottwald, R. L. (1971). Effects of response labels in concept attainment. *Journal of Experimental Psychology 91*(1), 30–33.

Greenberg, J. H. (1978). Diachrony, synchrony, and language universals. In J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (Eds.), *Universals of human language, volume 1, method and theory*, pp. 61–91. Stanford, California: Stanford University Press.

Griffiths, T. L. and M. L. Kalish (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science 31*(3), 441–480.

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

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

Hayes, B. (2009). *Introductory phonology*. Blackwell Textbooks in Linguistics. Blackwell.

Hayes, B. and C. Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning.

Linguistic Inquiry 39(3), 379–440.

Hayes, B., K. Zuraw, P. Siptár, and Z. Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language 85*(4), 822–863.

Haygood, R. C. and L. E. Bourne (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review 72*(3), 175–195.

Hunt, E. B. and J. M. Kreuter (1962, December). The development of decision trees in concept learning: III, learning the logical connectives. Working Paper 92, University of California, Los Angeles, Western Management Science Institute.

Hyman, L. M. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic studies offered to Joseph Greenberg: second volume: phonology*, pp. 407–418. Saratoga, California: Anma Libri.

Iverson, G. K. and J. C. Salmons (2011). Final devoicing and final laryngeal neutralization. In M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice (Eds.), *The Blackwell companion to phonology*, Chapter 69, pp. xx–xx. Blackwell.

Jackendoff, R. and S. Pinker (2005). The nature of the language faculty and its implications for the evolution of language (reply to Fitch, Hauser, and Chomsky. *Cognition 97*, 211–225.

Jäger, G. (2007). Maximum Entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (Eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan*, pp. 467–479. Stanford, California: CSLI Publications.

Jäkel, F., B. Schölkopf, and F. A. Wichmann (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences 13*(9).

Johnson, M. (2013, July). Language acquisition as statistical inference. Talk presented at the 19th International Congress of Linguists, Geneva.

Jurafsky, D. and J. H. Martin (2008). *Speech and Language Processing* (2nd ed.). Pearson Prentice Hall.

Kager, R. and J. Pater (2012). Phonotactics as phonology: Knowledge of a complex restriction in dutch. *Phonology 29*, 81–211.

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

Karev, G. P. (2010). Replicator equations and the principle of minimal production of information. *Bulletin of Mathematical Biology 72*, 1124–1142.

Kepros, P. C. and L. E. Bourne (1966). Identification of biconditional concepts: effect of number of relevant and irrelevant dimensions. *Canadian Journal of Psychology/Revue Canadienne de Psychologie 20*(2), 198–207.

King, W. L. (1966). Learning and utilization of conjunctive and disjunctive classification rules: a develop-

mental study. *Journal of Experimental Child Psychology 4*(3), 217–231.

King, W. L. and J. R. Holt (1970). Conjunctive and disjunctive rule learning as a function of age and forced verbalization. *Journal of Experimental Child Psychology 10*(1), 100–111.

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

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

Kruschke, J. K. (2005). Category learning. In K. Lamberts and R. L. Goldstone (Eds.), *The handbook of cognition*, Chapter 7, pp. 183–201. London: Sage.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*, Chapter 9, pp. 267–301. New York: Cambridge University Press.

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

Kurtz, K. J., K. R. Levering, R. D. Stanton, and J. R. anď Steven N. Morris (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*(2), 552–572.

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

Lai, Y. R. (2012). *Domain specificity in learning phonology*. Ph. D. thesis, University of Delaware.

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

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

Lewandowsky, S. (2011). Working memory capacity and categorization: individual differences and modelling. *Journal of Experimental Psychology: Learning, Memory, and Cognition 37*(3), 720–738.

Lewandowsky, S., B. R. Newell, L. Xieng Yang, and M. L. Kalish (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition 38*(4), 881–904.

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and*

Review 9(4), 829–835.

Love, B. C. and A. B. Markman (2003). The nonindependence of stimulus properties in human category learning. *Memory and Cognition 31*(5), 790–799.

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

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

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

Marcus, G. F., S. Vijayan, S. B. Rao, and P. M. Vishton (1999). Rule learning by seven-month-old infants. *Science 283*, 77–80.

Markman, A. B. and B. H. Ross (2003). Category use and category learning. *Psychological Bulletin 129*(4), 592–613.

Medin, D. L. and P. J. Schwanenflugel (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory 7*(5), 355–368.

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

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

Miller, R. R., R. C. Barnet, and N. J. Grahame (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin 117*(3), 363–386.

Minda, J. P., A. S. Desroches, and B. A. Church (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition 34*(6), 1518–1533.

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

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

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

Moreton, E. and J. Pater (2012a). Structure and substance in artificial-phonology learning: Part i, structure. *Language and Linguistics Compass 6*(11), 686–701.

Moreton, E. and J. Pater (2012b). Structure and substance in artificial-phonology learning: Part ii, substance. *Language and Linguistics Compass 6*(11), 702–718.

Moreton, E. and K. Pertsova (2012, October). Pastry phonotactics: Is phonological learning special? Presentation at the 43rd Annual Meeting of the Northeast Linguistic Society, City University of New York.

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

Newell, B. R., J. C. Dunn, and M. Kalish (2011). Systems of category learning: Fact or fantasy? In B. H. Ross (Ed.), *Psychology of learning and motivation: Advances in Research and Theory*, Volume 54, pp. 167–215. Academic Press.

Newport, E. (2011). The modularity issue in language acquisition: A rapprochement? Comments on Gallistel and Chomsky. *Language Learning and Development 7*(4), 279–286.

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

Nosofsky, R. M. and T. J. Palmeri (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review 3*(2), 222–226.

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

Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical linguistics: problems and perspectives*, pp. 237–278. Harlow: Longman.

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

Pater, J. and E. Moreton (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad 3*(2), 1–44.

Pertsova, K. (2012). Logical complexity in morphological learning. To appear in *Proceedings of the Berkeley Linguistics Society*.

Peters, K. G. and J. P. Denny (1971). Labelling and memory effects on categorizing and hypothesizing behavior for biconditional and conditional conceptual rules. *Journal of Experimental Psychology 87*(2), 229–233.

Pevtzow, R. and R. L. Goldstone (1994). Categorization and the parsing of objects. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Hillsdale, New Jersey, pp. 717–722. Lawrence Erlbaum Associates.

Prince, A. and P. Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Department of Linguistics, Rutgers University.

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

Pycha, A., E. Shin, and R. Shosted (2007). Directionality of assimilation in consonant clusters: an experimental approach. MS, Department of Linguistics, University of California, Berkeley.

Rafferty, A. N., T. L. Griffiths, and M. Ettlinger (2012). Greater learnability is not suffi-

cient to produce cultural universals. MS, Computer Science Division, Stanford University. URL: http://cocosci.berkeley.edu/tom/papers/RaffertyLearnabilityAndUniversals.pdf.

Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black and W. F. Prokasy (Eds.), *Classical conditioning*, Volume II: Current research and theory. New York: Appleton–Century–Crofts.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review 65*(6), 386–408.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*(5), 1249–1265.

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

Schane, S. A., B. Tranel, and H. Lane (1974). On the psychological reality of a natural rule of syllable structure. *Cognition 3*(4), 351–358.

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

Skoruppa, K. and S. Peperkamp (2011). Adaptation to novel accents: feature-based learning of context-sensitive phonological regularities. *Cognitive Science 35*, 348–366.

Smith, D. J. and J. P. Minda (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition 24*(6), 1411–1436.

Smith, J. D., M. E. Berg, R. G. Cook, M. S. Murphy, M. J. Crossley, J. Boomer, B. Spiering, M. J. Beran, B. A. Church, F. G. Ashby, and R. C. Grace (2012). Implicit and explicit categorization: a tale of four species. *Neuroscience and Biobehavioral Reviews xx*(xx), xxxx–xxxx.

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

Smith, J. D., J. I. Tracy, and M. J. Murray (1993). Depression and category learning. *Journal of Experimental Psychology: General 122*(3), 331.

Smolensky, P. and G. Legendre (2006). *The harmonic mind*. Cambridge, Massachusetts: MIT Press.

Staubs, R. and J. Pater (2014). Learning serial constraint-based grammars. In J. McCarthy and J. Pater (Eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Sutton, R. S. and A. G. Barto (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review 88*(2), 135–170.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model.

Cognition 92(1), 231–270.

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

Wilson, C. (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science 30*(5), 945–982.

Wong, P. C. M., M. Ettlinger, and J. Zheng (2013). Linguistic grammar learning and *DRD2*-taq-ia polymorp hism. *PLOS One 8*(5), e64983.

Yamauchi, T., B. C. Love, and A. B. Markman (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Co gnition 28*(3), 585–593.

Yamauchi, T. and A. B. Markman (1998). Category-learning by inference and classification. *Journal of Memory and Language 39*, 124–148.