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 October 6, 2013.

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

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

We test IMECCS, an implementation of Gluck and Bower (1988a)’s Configural Cue Model in 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 IMECCS 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.

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

<table>
<thead>
<tr>
<th>a. Phonological segments</th>
<th>b. Visual objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Voice</td>
</tr>
<tr>
<td>Coronal</td>
<td>[+voice]</td>
</tr>
<tr>
<td>Dorsal</td>
<td>[−voice]</td>
</tr>
</tbody>
</table>

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. Fig. 1 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.

First, 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 the distinction between “cue-based” and “rule-based” models (discussed at greater length in Section 5.4). Cue-based models learn by gradually updating the weights on a set of constraints (i.e., cue detectors), and their inductive biases arise from the constitution of the constraint set and the nature of the weight-update algorithm. We use the term “rule-based models” in a very narrow and specific sense, to refer to those models which learn by testing rule-like hypotheses and whose inductive bias respecting a hypothesis depends on the syntactic complexity of the hypothesis (e.g., as a Boolean formula). 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 extant phonological 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 IMECCS (Incremental 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, IMECCS can be seen as an incremental 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 IMECCS 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 supervised phonotactic learning of Types I, II, and IV (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, IMMECS 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 IMMECS, they argue for a reconsideration of cue-based learning modes. 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).

Finally, the joint study of phonological and visual concept learning 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.) A satisfactory answer to this question with respect to concept learning will require new studies that control task and stimulus factors. The present results point to the availability of similar cue-based learning processes for analogous phonotactic and visual patterns. 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 learning. 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 et al. (2013) for a recent review).

Models of concept learning fall into three basic categories (Kruschke, 2005; Ashby and Maddox, 2005; Kruschke, 2008). Rule-based models learn by testing hypotheses that are stated as logical combinations of stimulus features (see recently Goodman et al. 2008, Pertsova 2012a). Cue-based models, such as the single-layer perceptron, learn by gradually updating weights on a pre-specified set of property detectors (see recently Kurtz 2007). Exemplar-based models classify stimuli on the basis of similarity to previously-seen exemplars, and learn by both memorizing exemplars and adjusting the parameters of the similarity function. Many variations and hybrids are possible, but rules, cues, and exemplars remain the central ideas. As we will see in detail in section 3, pure cue-based models have been unsuccessful in matching the classical Shepard 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 by using inductive bias that makes them less sensitive to patterns that crucially involve more features. Rule-, cue- and exemplar-based models have been proposed and compared in phonology (e.g. Chomsky and Halle 1968; Rumelhart and McClelland 1986; Pierrehumbert 2001), but to our knowledge, no prior research has examined their predictions for learning of phonological instantiations of the Shepard types.1

---

1Chomsky and Halle (1968, p. 334) propose an evaluation procedure that prefers grammars that use fewer features, but it was originally proposed only as a means of choosing amongst analyses of a single set of data, and does not make one pattern easier to learn than another.
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; Kurtz et al., 2013; Moreton, 2012).

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 structurally simple pattern, using a number of vowel and consonant features, with a wide range of methodologies, and both infant and adult participants.

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. The training methodology used in this study involved a combination of unsupervised 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 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, also with the proviso that the trained concept space had only 6 members (testing involved generalization to a novel sound).

A Type I advantage was also found using supervised learning 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 tʰ p pʰ], 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 pʰj tw tʰw] as IN) to Type VI (e.g. [pj tʰj tw pʰw] 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 are only given positive examples. Phonological alternations, changes in the shapes 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 indicates that it is not [z] or [ç] in that context. Studies of the learning of alternations have therefore often been run using supervised learning, 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 [-ck] or [-ak] 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). 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. 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, 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. Agreement, 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 might be that it has been assumed to be trivial to model the Type I single feature advantage, or a Type II over Type VI advantage, which is essentially a difference between there being a rule or not. A survey of the learning models proposed for visual concepts shows that this assumption would in fact be correct: all models that we

---

2There 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.
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 that has no explicitly stated preference for patterns with fewer features, and which in fact also yields $III, IV > II$, which contradicts a preference for fewer features.

3 The IMECCS model

This section describes IMECCS (Incremental 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. The model can be viewed as a MaxEnt version of a well-known psychological model of concept learning, the Configural Cue Model of Gluck and Bower (1988a,b). Alternatively, it can be seen as a version of the phonotactic learner of Hayes and Wilson (2008), which uses unbiased conjunctive constraints rather than inducing them from the training data. 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 the criterion by which models of human category learning were judged, and the standard to which they were engineered. 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. IMECCS is an adaptation of the Configural Cue Model to current phonological learning paradigms.
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, IMECCS, follows Hayes and Wilson in using MaxEnt to state a probability distribution over the space of possible word forms, and thus as 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, though we use batch, rather than stochastic gradient descent.

The target distribution of stimuli is given by a random variable $X$ which takes on values $x_1, \ldots, x_n$ with probabilities $\Pr(x_1), \ldots, \Pr(x_n)$. 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 the model’s estimate of $\Pr(X = x_j | \mathbf{w})$.

The assignment of probabilities to word forms is illustrated in Fig. 3 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, $+vce \wedge \text{Cor}$ targets the one consonant that is both voiced and coronal, $[d]$, and $-vce \wedge \text{Dor}$ 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:

\[
\hat{\Pr}(x_j | \mathbf{w}) = \frac{e^{h_{\mathbf{w}}(x_j)}}{\sum_{x \in D} e^{h_{\mathbf{w}}(x)}}
\]
\[ h_w(x_j) = \sum_{i=1}^{m} w_i c_i(x_j) \]  

(1)

\[ Z_w = \sum_{j=1}^{n} \exp h_w(x_j) \]  

(2)

\[ \Pr(X = x_j \mid w) = \frac{\exp h_w(x_j)}{Z_w} \]  

(3)

In the example of Fig. 3, 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 \( p \).

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

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

(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 \( \eta \):

\[ \Delta w_i = \eta \cdot (E_{emp}[c_i(X)] - E_w[c_i(X)]) \]  

(5)

This model is an unsupervised 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 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.,

---

3The 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 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.
to use gradient descent rather than stochastic gradient descent) because the batch learner conveniently and
tractably approximates an average over many independent runs of an on-line simulation.\(^4\) 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,\(^5\) 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.\(^6\) The updates
can be done directly on the model’s estimated probabilities using the following update rule:

\[ \Delta p_j = \eta \cdot p_j \cdot \sum_{i=1}^{n} (c_i(x_j) - q_i) \cdot (q^* - q_k) \]  

(6)

where \( q_i \) and \( q_k \) are the model’s expectation of \( c_i \) and \( c_k \), and \( q^* \) is the true mean of \( c_k \) in the target
distribution. This result is advantageous for two 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 other, discussed at greater length in Section 4.1, is that the factors of the summand have intuitive
interpretations in terms of constraint properties.

### 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 \( \alpha \) can have the values +, –, or ± (i.e.,
“don’t care”). A conjunctive constraint of order \( k \) on \( N \) features is one that has exactly \( k \) coefficients that
are not ±. (In Fig. 3, the constraint [+voice] is a conjunctive constraint of order 1, while the constraint
 [+voice] \( \land \) [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

---

\(^4\)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.

\(^5\)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 IMECCS
emerge from the structure of the constraint set.

\(^6\)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.
conjunctive constraint sets.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Order</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\pm F_1, \pm F_2)$</td>
<td>0</td>
<td>Always 1</td>
</tr>
<tr>
<td>$(+F_1, +F_2)$</td>
<td>1</td>
<td>1 iff stimulus is $+F_1$</td>
</tr>
<tr>
<td>$(-F_1, +F_2)$</td>
<td>1</td>
<td>1 iff stimulus is $-F_1$</td>
</tr>
<tr>
<td>$(\pm F_1, +F_2)$</td>
<td>1</td>
<td>1 iff stimulus is $+F_2$</td>
</tr>
<tr>
<td>$(\pm F_1, -F_2)$</td>
<td>1</td>
<td>1 iff stimulus is $-F_2$</td>
</tr>
<tr>
<td>$(+F_1, +F_2)$</td>
<td>2</td>
<td>1 iff stimulus is simultaneously $+F_1$ and $+F_2$</td>
</tr>
<tr>
<td>$(-F_1, +F_2)$</td>
<td>2</td>
<td>1 iff stimulus is simultaneously $-F_1$ and $+F_2$</td>
</tr>
<tr>
<td>$(+F_1, -F_2)$</td>
<td>2</td>
<td>1 iff stimulus is simultaneously $+F_1$ and $-F_2$</td>
</tr>
<tr>
<td>$(-F_1, +F_2)$</td>
<td>2</td>
<td>1 iff stimulus is simultaneously $-F_1$ and $-F_2$</td>
</tr>
</tbody>
</table>

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

Conjunctive constraints have been previously proposed in the Optimality Theory literature, where they 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 IMECCS includes all of them, making no attempt to winnow out 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).

A useful property of the complete conjunctive constraint set is that every stimulus $x_j$ has a characteristic constraint that is 1 for $x_j$ and 0 for all other stimuli. For convenience, we assume that the constraints are numbered so that the characteristic constraint for $x_j$ is $c_j$. The model’s estimate of the probability of $x_j$ is then just its expectation of $c_j(X)$, and the rate at which $x_j$ is gaining probability is just $\Delta p_j$ in Equation 6.

3.3 Relationship to previous domain-general learning models

We have just introduced IMECCS 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. IMECCS is simply the Configural Cue Model in the Maximum Entropy framework.

4 IMECCS 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 IMECCS. 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 IMECCS 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. We will then show that the same behavior emerges in an 8-feature space like the one used in our own experiments.

4.1 Learning the SHJ patterns in a three-dimensional space

Figure 4 shows learning curves for all six types. Each epoch is a single update using the entire probability distribution as described in Section 3. The vertical axis shows the summed probability of the positive stimuli. 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. Probability shifts 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 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 IMECCS as it does in the Configural Cue Model.
Figure 4: Learning curves for Shepard Types produced by MaxEnt with Gradient Descent
The relative quickness with which each of the Types is learned is a function of the structure of the constraint set. The way that the constraints apply to each of the Types is illustrated in the cubes of figure 5, adopted from Shepard et al. (1961). The cubes show the concepts in a three-dimensional space corresponding to the three features, with the IN class indicated with black dots on the vertices. In these diagrams, the top and bottom faces of the cubes correspond to black and white in the shapes below, left and right to circle and triangle, and front and back to small and big.

The unbiased conjunctive constraint schema provides a constraint for each corner, edge, and face on the cube, as 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. The reason that the SHJ Types are learned at different rates is that they differ in the degree to which the stimuli share edge and face constraints with each other.

For example, consider the positive stimuli of Type IV at the very outset of learning. Type IV has a core-periphery structure with one central stimulus $x_8$ surrounded by three peripheral stimuli $x_4$, $x_6$, and $x_7$. The central stimulus is directly supported by its own corner constraint, which is a valid positive constraint because it gives a 1 only to positive stimuli ($x_8$ and no others), by three valid positive edge constraints, and by three partially-valid positive face constraints. By contrast, a peripheral stimulus like $x_7$ is directly supported by one valid corner constraint, one valid edge constraint, and two partially-valid face constraints.\(^7\)

---

\(^7\)By valid constraint, we mean a constraint that is reliably informative, i.e., one that picks out a subset of the positive stimuli or a subset of the negative stimuli. 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. 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. An invalid constraint is one whose value is 1 on equally many positive and negative stimuli.
Table 2: The conjunctive constraints and their contribution to the 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 \(c_k\). Black and white dots • and ○ represent positive and negative stimuli respectively, and \(x_j\) is marked as •. Rotations and reflections are omitted. Columns: \(q^*_k\) is the true expectation of \(c_k\) in the target distribution; \(q_k\) is the model’s current expectation; \(c_k(x_j)\) is the value of Constraint \(c_k\) on Stimulus \(x_j\). Not shown is the bias constraint, whose value is 1 for all 8 stimuli.
A constraint $c_k$ supports a stimulus $x_j$ by its contribution $(q^*_k - q_k) \cdot (c_k(x_j) - q_k)$ to the learning rate $\Delta p_j$ in Equation 6. Valid constraints are effective supporters because they are right more often than expected by chance, and hence gain weight quickly. This shows up in the factor $(q^*_k - q_k)$, which measures the difference between the actual and the expected frequency with which training stimuli (all positive) get a score of 1 from $c_k$. This difference is doubled for a valid edge compared to a valid corner, because the edge constraint applies to twice as many stimuli, and quadrupled for a valid face; thus, when validity is equated, more-general constraints provide more support.

The other factor determining the contribution of $c_k$ is $(c_k(x_j) - q_k)$, which measures how atypical $x_j$ is on constraint $c_k$, i.e., how specific $c_k$ is to $x_j$. Among positive constraints in the IMECCS constraint set, 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.

The product of these factors, shown in the last column of Table 2, is the contribution of each kind of constraint in early learning. Positive constraints contribute in the descending order: valid faces, valid edges, partially-valid faces, and valid corners. Of these, the most influential 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 earliest, and those that are isolated (no positive edge neighbors) gain it latest. This is borne out by the simulations shown in Figure 6; 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 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.

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 IMECCS 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 distribution assigned 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):

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

(7)

The resulting generalization-performance curves, averaged over 30 simulations in each SHJ Type, are shown in Figure 7. They are qualitatively very similar to the learning curves for the three-feature model shown in Figure 4. Performance is always best in Type I. The next best are Types III and IV. Early on,
Figure 6: Probability allocated to a stimulus in each of the different positive cells within each SHJ Type as a function of time. The learning target is \( \Pr(x) = \frac{1}{4} \). (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 cell 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. (Learning rate \( \eta = 0.01 \).)
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 $I > III > IV > II > V > VI$.\(^8\)

In sum, IMECCS has an inductive bias for learning of Shepard Types with an ordering $I > III, IV > HI, V > VI$. 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 IMECCS replicate the original Configural Cue Model’s prediction that Type II will overtake Type IV late in learning.\(^9\) IMECCS learning in IMECCS is faster with 8 features than with 3. This effect can be seen in Equation 6: Adding constraints increases $\eta$ without reducing any of the other variables, and so increases the size of the weight update $\Delta \beta_j$. 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 $\eta$ 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.
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 IMECCS, rather than as described by rule-based models of concept learning such as RULEX (Nosofsky et al., 1994) or the Rational Rules model (Goodman et al., 2008). Theoretically, IMECCS 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, by comparing 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 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 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).

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. Examples are shown in Table 4.
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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Stimulus position</th>
<th>Consonants</th>
<th>Vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1$</td>
<td>$V_1$</td>
<td>$C_2$</td>
</tr>
<tr>
<td>voiced</td>
<td>±</td>
<td>±</td>
<td></td>
</tr>
<tr>
<td>Coronal</td>
<td>±</td>
<td>±</td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>±</td>
<td>±</td>
<td></td>
</tr>
<tr>
<td>back</td>
<td>±</td>
<td>±</td>
<td></td>
</tr>
</tbody>
</table>

L1 (TYPE I): $C_1$ is voiced
digu, gada, dika, gugu, …

L2 (TYPE II): $C_1$ is voiced iff $V_2$ is back.
digu, taŋi, kagaæ gada, …

L3 (TYPE IV): At least two of: $C_1$ is voiced, $V_2$ is high, $V_2$ is back
kaku, digu, guki, dækæ, …

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

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 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.

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, 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 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, IMECCS predicts generalization performance to follow the order \( I > III, IV > II, V > VI \). Rule-based models engineered to 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).

5.3 Results

Fig 8 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 \( I > IV > III > V > II > VI \). This

\(^{10}\)This criterion was chosen because it is significantly below chance at the 95% level by an exact binomial test.
Figure 8: Individual participant performance (proportion pattern-conforming responses) for Experiment 1. 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.

differs from the classic SHJ order $I > II > III, IV, V > VI$, but matches the $I > III, IV > II, V > VI$ order predicted by IMECCS.

<table>
<thead>
<tr>
<th>Type</th>
<th>I</th>
<th>II</th>
<th>II</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plotting symbol</td>
<td>◦</td>
<td>◦</td>
<td>△</td>
<td>+</td>
<td>◦</td>
<td>◦</td>
<td>◦</td>
<td>◦</td>
</tr>
<tr>
<td>Mean</td>
<td>0.737</td>
<td>0.567</td>
<td>0.598</td>
<td>0.708</td>
<td>0.668</td>
<td>0.704</td>
<td>0.651</td>
<td>0.594</td>
</tr>
<tr>
<td>SD</td>
<td>0.120</td>
<td>0.112</td>
<td>0.123</td>
<td>0.153</td>
<td>0.129</td>
<td>0.091</td>
<td>0.091</td>
<td>0.099</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>sem</td>
<td>0.024</td>
<td>0.032</td>
<td>0.050</td>
<td>0.062</td>
<td>0.026</td>
<td>0.018</td>
<td>0.018</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Table 5: Mean proportion correct in each Type condition of Experiment 1. Plotting symbols refer to Fig. 8.

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 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 mentioned in Section 5.1.1 had no effect in Types III—VI, so these sub-conditions were collapsed together within each of these Type conditions.\textsuperscript{11} They were retained for Type

\textsuperscript{11}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 determine how much of Type I’s advantage comes from using only one feature, and how much comes from not mixing segments or genera.
II, and renamed to indicate their restriction to Type II. A Type II pattern for which \( \text{Both Same Genus} = 1 \) is a harmony or disharmony pattern (e.g., \( V_1 \) is [+high] iff \( V_2 \) is [−high]), and there is independent evidence that such patterns can be easier to learn than other Type II patterns (see review in Moreton and Pater 2012a, §3.1). A Type II pattern for which \( \text{Both Same Segment} = 1 \) can be solved by attending to a single segmental position, which might make the pattern easier to detect.

Two nuisance factors, \( \text{Redup} \) and \( \text{CorrFirst} \), 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. \( \text{Redup} \) 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. \( \text{CorrFirst} \) 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 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, \( \text{Both Same Genus} \) did not differ significantly from zero. The significant positive estimate for \( \text{Both Same Segment} \) 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 \( \text{CorrFirst} \) and \( \text{Redup} \) had significant effects, indicating that participants were biased to choose the first 2AFC response and to avoid reduplicated words.

| Coefficient               | Estimate | SE     | z     | Pr(>|z|)  |
|---------------------------|----------|--------|-------|----------|
| \((\text{Intercept})\)   | 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 |
| \(\text{Both Same Genus}\)| 0.11078  | 0.24329| 0.455 | 0.648850 |
| \(\text{Both Same Segment}\) | 0.69318 | 0.25149| 2.756 | 0.005845 |
| \(\text{CorrFirst}\)     | 0.27396  | 0.06348| 4.316 | <0.0001  |
| \(\text{Redup}\)         | -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 based on those experiments, Type II is not easier than Types III, IV, and V, but harder, as predicted by IMECCS. IMECCS has only one free parameter, the learning rate \( \eta \) (in our simulations, always 0.01). Changing \( \eta \) changes only the time scale, and is equivalent to multiplying the trial numbers
by a constant factor. To find the best IMECCS fit to the human data, we found it convenient to leave the time scale as it was and instead find the trial that minimized the disparity between human and IMECCS 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. By using the modelled cell means rather than the raw ones, we removed the effects of the nuisance variables (which IMECCS is not designed to account for). Disparity at a given trials was taken to be the squared difference between these cell means and the IMECCS proportion correct at that trial. The best match (disparity < 0.0032) was attained on Trial 9, as shown in Fig. 9. The relative numerical order, I > IV > III > V > II > VI, was in agreement with the IMECCS predictions (Section 5.2). Performance on Type II was lower than the IMECCS 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 IMECCS 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 9: Human performance most closely matches IMECCS predictions at Trial 9. (The Type II mean excludes the same-segment and same-feature-genus cases.) Curves show mean of 100 replications each; \( \eta = 0.01 \). 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 IMECCS (see previous discussion in Section 4.1). The probability that IMECCS 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). Fig. 10 shows 2AFC performance by humans and by IMECCS on Trial 9. 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 IMECCS (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 10: Effect of edge constraints in human learners (left panel) and in IMECCS (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. $\text{IIss} = \text{Type II with both features in the same segment}$ (humans only).

5.4 Discussion

Where the predictions of IMECCS differed from the classic SHJ order, participants in Experiment 1 followed the IMECCS predictions. The resemblance to IMECCS 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 IMECCS in a detailed way, and support the use of constraint-based phonological models.
more generally. We now turn to the question of whether they are also consistent with rule-based models. (We remind the reader 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 IMECCS (Fig. 10).

The Rational Rules model (Goodman et al., 2008) generates hypotheses using a grammar of logical formulas, to which it then assigns prior probabilities according to their syntactic structures. It 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).

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 learning and explicit rule-seeking are known to facilitate Type II relative to Type IV in non-linguistic learning (Love, 2002; Love and Markman, 2003; Lewandowsky, 2011; 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 IMECCS, see Footnote 8, 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 verbal 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Stimulus segment</th>
<th>Nonlinguistic analogues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1$</td>
<td>$\sigma_2$</td>
</tr>
<tr>
<td>voiced</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>Coronal</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>high</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>back</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. (1) Each word stimulus from Experiment 1 should have a unique visual analogue in Experiment 2. (2) 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. (3) 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. (4) The objects should be relatively “natural” in the sense of being of a recognizable and familiar sort.

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. 11.

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”.

31
Figure 11: Examples of corresponding cakes and words.

Figure 12: Sequence of events for the familiarization phase of Experiment 2.

[figure showing the sequence of events with icons]

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.
6.2 Results and discussion

Individual participant means are shown in Table 8 and in Fig. 13. 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
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).

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

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

| 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 |

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, which would afford 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 patterns did not afford verbally simple rules and hence did not facilitate explicit rule-based learning. 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.\(^{12}\)

Extraction of rules, whether explicit or implicit, may also be inhibited by perceptual integrality between stimulus dimensions, which prevents selective attention to individual features. When Nosofsky and Palmeri (1996) 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. Still, it is not likely that perceptual integrality is the cause of the high relative difficulty of Type II in Experiments 1 and 2. There is evidence of some 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), but 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.

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 in both domains. This is confirmed when we examine the effect of edge constraints on two-alternative forced-choice responses, shown in Fig. 14. 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. 10). The results of Experiments 1 and 2 together therefore support the hypothesis that both a rule-based system and a cue-based system are available to the learner, and can be applied to both phonotactic and visual patterns (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al., 2012).

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 of the same paradigm evokes cue-based learning in some conditions, but rule-based learning in others, thereby confirming that the cue-based performance in Experiment 1 is not an inevitable consequence of the paradigm. Two

\(^{12}\)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) and Wong et al. (2013) in morphology.
Figure 14: 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.
main possibilities remain. One is that all phonotactic learning takes place via cue-based 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. (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 authors’ 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.
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 (see Appendix for instructions). 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. 15. 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.

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. 15.

<table>
<thead>
<tr>
<th>Type</th>
<th>I</th>
<th>II</th>
<th>II</th>
<th>II</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plotting symbol</td>
<td>◦</td>
<td>○</td>
<td>△</td>
<td>+</td>
<td>◦</td>
</tr>
<tr>
<td>Mean</td>
<td>0.688</td>
<td>0.479</td>
<td>0.651</td>
<td>0.578</td>
<td>0.570</td>
</tr>
<tr>
<td>SD</td>
<td>0.203</td>
<td>0.077</td>
<td>0.163</td>
<td>0.224</td>
<td>0.129</td>
</tr>
<tr>
<td>N</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>s.e.m.</td>
<td>0.083</td>
<td>0.031</td>
<td>0.067</td>
<td>0.091</td>
<td>0.026</td>
</tr>
</tbody>
</table>

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 15: 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 $\times$ 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 $\times$ I | 0.315515 | 0.283824 | 1.112 | 0.26629 |
| Supervised $\times$ IV | −0.154627 | 0.332198 | 0.471 | 0.63516 |
| Supervised $\times$ Both Same Segment | 0.004564 | 0.332198 | 0.014 | 0.98906 |
| Supervised $\times$ 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 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. 16. 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. 16 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 (the Block $\times$ IV term was positive, but not significantly so). The significant Block $\times$ 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 IMECCS. IMECCS 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.\(^{13}\) As Fig. 17

\(^{13}\)This has the effect of treating a classification task as a stimulus-completion or inference task. IMECCS 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 presumabley find an even larger advantage for Type IV over Type II. Thus, even if the nine-feature IMECCS is actually a more appropriate model of inference than of classification, it is still supported (albeit more weakly) by the results of Experiment 3.
| 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 | 0.556 | 0.5919   |
| Block                       | −0.03165 | 0.06479 | −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. 7).

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 visual pattern learning have tended to focus structural effects on supervised learning 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, and extending them to all six SHJ pattern types. Participant behavior in both domains was well described by IMECCS, a cue-based learner which fuses the (visual) Configural Cue Model (Gluck and
Figure 17: Predicted generalization performance (probability of correctly classifying a novel stimulus) for 8-feature IMECCS under supervised training (average of 250 replications per Type; learning rate $\eta = 0.01$). Chance performance is 0.5.
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).

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 and instead derives categorization behavior from similarity with stored exemplars. (For a recent review of exemplar models in psychology, see Kruschke 2008; in phonology, Kirchner et al. 2010). Abstraction in IMECCS 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 IMECCS are the estimated probabilities of the individual stimuli, which raises the question of whether IMECCS (and other Max Ent models described by Equations 1–??) is an exemplar model in disguise. This possibility is reinforced by the observation that the predictions of IMECCS 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.\textsuperscript{14}

However, there are also major differences between the weightless IMECCS and exemplar models. Most notably, the IMECCS 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 IMECCS 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 IMECCS 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.

\textsuperscript{14}In the vectorized weightless 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 IMECCS, 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 IMECCS 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 \(C\) measures featural similarity between stimuli.
It is a separate question whether the outward behavior of IMECCS, or of the humans in these experiments, can be replicated by an exemplar model with an appropriate similarity function. 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 learning 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 Incremental 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 analyses 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 Incremental 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 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. IMECCS takes an intermediate approach by pre-loading the constraint

---

15Because the similarity matrix $C^T C$ 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 IMECCS and related models.
set with all constraints that conform to the conjunctive schema.\textsuperscript{16} This choice makes IMECCS 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).\textsuperscript{17} 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 and 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 languages; e.g., many languages require syllable- or word-final obstruent consonants to lack voicing, but few or none require them to possess it (Iverson and Salmons, 2011). However, an intensive search for learnability differences answering to these typological asymmetries has, when examined closely, so far found only weak and inconsistent corroboration (reviewed in Moreton and Pater 2012b).

When instead the features are kept constant and the pattern structure is manipulated, the effects on

\textsuperscript{16}IMECCS 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).

\textsuperscript{17}The authors are indebted to [name omitted] for pointing out the relevance of this literature.
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, 2012b). 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 without weights

This appendix shows that any Maximum Entropy learner that uses (unregularized) gradient descent on negative log-likelihood can be converted into an equivalent model in which the stimulus probabilities, rather than the weights, are the state variables. The equivalence is not limited to the IMECCS learner described in the text. We will use the following notation: there are \( m \) constraints (“features” in machine-learning terms) \( \{c_i\}_{i=1}^m \), \( m \) weights \( \{w_i\}_{i=1}^m \), and \( n \) stimuli \( \{x_j\}_{j=1}^n \). The empirical expectation of a random variable \( X \) is denoted \( \text{E}_{\text{emp}}[X] \), while the learner’s expectation when the weight vector \( w = (w_1, \ldots, w_m) \) is \( \text{E}_w[X] \). 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 \( x_i \):

\[
\frac{\partial}{\partial w_k} \text{E}_w[c_i] = \frac{\partial}{\partial w_k} \sum_{j=1}^n \text{Pr}(x_j | w) \cdot c_i(x_j) \\
= \sum_{j=1}^n \frac{\partial}{\partial w_k} \text{Pr}(x_j | w) \cdot c_i(x_j) \\
= \sum_{j=1}^n \left( c_i(x_j) \cdot \frac{\partial}{\partial w_k} \text{Pr}(x_j | w) \right)
\] (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:

\[
\frac{\partial}{\partial w_i} \text{Pr}(x_j | w) = \frac{\partial}{\partial w_i} \frac{\exp h_w(x_j)}{Z_w} = \frac{\exp h_w(x_j)}{Z_w} \cdot \frac{\partial}{\partial w_i} \frac{\exp h_w(x_j)}{Z_w} = \text{Pr}(x_j | w) \cdot \frac{\partial}{\partial w_i} \frac{\exp h_w(x_j)}{Z_w}
\] (9)

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

\[
\frac{\partial}{\partial w_i} \text{Pr}(x_j | w) = \frac{\exp h_w(x_j)}{Z_w} (c_i(x_j) - E_w[c_i])
\] (10)

Substituting Equation 10 into Equation 8 now yields:
The Gradient Ascent 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

\[
\frac{d}{dt} E_w[c_i] = \sum_{k=1}^{n} \frac{\partial}{\partial w_k} E_w[c_i] \cdot \frac{\partial w_k}{\partial t} = \eta \sum_{k=1}^{n} \text{Cov}_w[c_i, c_k] \cdot (E_{\text{emp}}[c_k] - E_w[c_k])
\]

Now we augment the constraint set by adding, for every stimulus \(x_j\), a “characteristic constraint” \(\hat{c}_j\) whose value is some small \(\epsilon\) for \(x_j\) and 0 for any other stimulus. The characteristic constraints have three useful properties. The first is that by choosing \(\epsilon\) 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., \(E_w[\hat{c}_j] = \epsilon \text{Pr}(x_j | w)\) (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, \(E_w[c_i\hat{c}_j] = c_i(x_j) \cdot E_w[\hat{c}_j]\). The covariance between a characteristic constraint and any other constraint is therefore

\[
\text{Cov}_w[c_i, \hat{c}_j] = E_w[c_i, \hat{c}_j] - E_w[c_i]E_w[\hat{c}_j]
\]

The covariance between two characteristic constraints is zero unless they are both the same constraint, in which case it is the negligibly small quantity \(\epsilon^2\), so we can safely ignore them. Now we can substitute
Equation 13 back into Equation 12 to get:

\[
\frac{d}{dt} \Pr(x_j | w) = \frac{d}{dt} \frac{1}{\epsilon} E_w[\hat{c}_j]
\]

\[
= \eta \sum_{i=1}^{n} \frac{1}{\epsilon} \text{Cov}_w[c_i, \hat{c}_j] \cdot (E_{emp}[c_i] - E_w[c_i])
\]

\[
= \eta \cdot \Pr(x_j | w) \cdot \left( \sum_{i=1}^{n} (c_i(x_j) - E_w[c_i]) \cdot (E_{emp}[c_i] - E_w[c_i]) \right)
\]

(14)

For brevity, let \( p_j = \Pr(x_j | w) \), let \( q_i = E_w[c_i] \), and let \( q^*_i = E_{emp}[c_i] \). Then Equation 14 becomes

\[
\frac{d}{dt} p_j = \eta \cdot p_j \cdot \left( \sum_{i=1}^{n} (c_i(x_j) - q_i) \cdot (q^*_i - q_k) \right)
\]

(15)

which completes the derivation of Equation 6. If the model’s current estimated probabilities \( \{p_j\}_{i=1}^{n} \) for all stimuli are known, then \( w \) is not needed to evaluate Equation 15, because we can calculate \( q_i \) and \( q_k \) by simply adding up the probability of each stimulus times its score on \( c_i \) or \( c_k \). To calculate \( q = \{q_i\}_{i=1}^{m} \) from \( p = \{p_j\}_{i=1}^{n} \), let \( C \) be the matrix whose \( j \)-th column is the score vector of \( x_j \), i.e., \( C_{i,j} = c_i(x_j) \). Then \( q = C p \).

If we let \( e = p^* - p \) be the model’s error, then Equation 15 yields the following update rule:

\[
\Delta p = \eta \cdot \text{diag}(p)[S e - 1_n(S e)^T p]
\]

(16)

where \( \text{diag}(p) \) is the \( n \times n \) matrix with \( p \) along the main diagonal and 0 elsewhere, \( S = C^T C \), and \( 1_n \) is a column vector of \( n \) ones. 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 in constraint-score space.

This derivation was tested by simulation. For each simulation run, \( C \) was a 27-by-8 matrix, like 3-feature IMECCS, with entries drawn from a standard normal distribution (mean 0, s.d. 1). The target distribution \( p^* \) was made by choosing 8 values from a uniform distribution and then normalizing by their sum. The original algorithm and the weightless algorithm were then run for 1000 trials each with a learning rate of \( \eta = 0.01 \). For each trial, two measures of discrepancy were computed: the RMS difference between the two estimates of \( p \), and the largest absolute difference between any two \( p_j \). The discrepancies for an entire run were taken to be the maxima of these two measures across all 1000 trials. This was done for 1000 simulation runs. The maximum absolute difference across all runs was 0.004386448; the maximum RMS difference, 0.001717374. The respective medians were 0.0002787986 and 0.00014572.
References


Coberly, M. S. and A. F. Healy (1984). Accessibility of place and manner features and the place/manner


In J. Spenader, A. Erkisson, and O. Dahl (Eds.), Proceedings of the Stockholm Workshop on Variation within Optimality Theory, pp. 111–120.


Smith, J. D., J. P. Minda, and D. A. Washburn (2004). Category learning in rhesus monkeys: a study of


