Modelling modularity bias in phonological pattern learning

Elliott Moreton
University of North Carolina, Chapel Hill

[1]

1. Introduction

Phonological dependencies in natural language tend to relate elements which are phonetically similar. As Yip (2002:222) puts it, "In general in phonology, like things interact. We do not expect to find labials causing vowel fronting, or laterals blocking rounding harmony, for example. What makes sounds 'alike' are their articulatory and acoustic properties phonologized as feature specifications." Two main causal factors have been implicated in this tendency.

One is *channel bias*, phonetically-systematic errors in transmission between speaker and hearer that mislead learners into acquiring a grammar different from the ambient one. Physical links between similar units may skew the errors, and, through them, the grammars (Ohala, 1994; Beddor et al., 2001; Barnes, 2002; Kavitskaya, 2002; Blevins, 2004). Quite a lot is known about channel biases related to phonetic similarity, such as effects of coarticulation, acoustic confusability, and motor or perceptual priming. There is no dispute about their existence, only about their constitution and their responsibility for natural-language typology (e.g., Blevins (2006); de Lacy (2006); Kiparsky (2006); Hansson (2008)).

The other factor is *analytic bias*, disparities in how easily or accurately different patterns can be learned from training data of equal statistical quality (Wilson, 2003a). The general principle that X–X dependencies are somehow privileged over X–Y dependencies is supported by laboratory evidence (Pycha et al., 2003; Wilson, 2003b; Newport & Aslin, 2004; Onnis et al., 2005; Moreton, 2008), and the literature contains several theoretical proposals along those lines:

- Analytic bias favors dependencies which involve one feature over those which involve two (Clements & Hume, 1995; Pycha et al., 2003; Gordon, 2004).
- Analytic bias favors dependencies on a single autosegmental or Feature-Geometric tier over those which cross tiers (Goldsmith, 1976; McCarthy, 1981; Clements, 1995; Newport & Aslin, 2004).
- Analytic bias favors dependencies between featurally-similar elements (Frisch et al., 2004; Rose & Walker, 2004; Onnis et al., 2005).

Some of these proposals address bias directly through constraints on the learning process, others indirectly through a theory of Universal Grammar. However, two main things are lacking, which this paper aims to supply: an explicit account of how the bias operates in a learner, and a motivation for the bias. It is unsatisfying to note that X–X dependencies are "more salient" and stop there; the learning model ought to *make* them more salient. The common way to do that is to hard-wire a heuristic, instructing the learner to penalize X–Y hypotheses by trying them later or giving them lower prior probability.

In the alternative presented here, bias emerges out of the learner's preference for good explanations. Parametrized grammar schemas compete to explain the training data. A "good explanation" is a schema which makes the training data relatively probable. Schemas with fewer adjustable parameters can fit fewer different training sets, but assign higher probability to those which they do fit—the "Bayesian

[1]The ideas in this paper have benefited from conversations with many people, including Adam Albright, Joe Pater, Jason Riggle, Jennifer Smith, Anne-Michelle Tessier, and audiences at the Phonologization Symposium at the University of Chicago and WCCFL 30 at UCLA. Any remaining errors are mine alone. Email may be addressed to moreton@unc.edu.

© 2008 Elliott Moreton
Cascadilla Proceedings Project
Completed August 2, 2008

Occam's Razor" (MacKay, 2003, 343ff)—hence, schemas with fewer parameters are preferred by the learner when consistent with the data. When an X–X dependency is parsed out of the stimulus, it leaves a featurally-symmetric residue which can be accounted for with fewer parameters than the asymmetric residue left by an X–Y dependency. The result is a learning advantage for X–X dependencies over X–Y ones. The principle is a general one that can be applied to a wide range of learning models to derive *modularity bias*, a preference for grammars which minimize interaction between phonological subsystems.

The rest of this paper is structured as follows. Section 2 reviews the human data to be modelled. Section 3 describes BLUMPS, a simple phonotactic learner which exhibits modularity bias. BLUMPS is used to simulate the human data in Section 4. Section 5 analyzes BLUMPS to confirm that it works for the intended reasons. Concluding discussion is in Section 6.

2. Empirical basis

This paper will focus on a specific instance of modularity bias, the superior learnability of a dependency between the height features of two vowels compared to one between the height of a vowel and the voicing of a following consonant. This case is of particular interest because it matches an asymmetry in the typological frequency of the two patterns in natural language, one which is not explained by a difference in the magnitude of the phonetic precursors Moreton (2008). The data to be modelled comes from a series phonotactic-pattern-learning experiments with adult English speakers. The experiments will be described in detail in a separate paper; four are briefly summarized here.

The experimental paradigm is a modification of that used in Moreton (2008). All of the experiments used stimuli drawn from the same set and participants drawn from the same population, as well as the same procedures and apparatus; the only differences were in how the stimuli were chosen. Experiment 1 will be described first as a representative; the others are simple variations on it.

Stimuli were $C_1V_1C_2V_2$ words with inventory /t k d ɡ/ /i u æ ɔ/, synthesized using the MBROLA concatenative diphone synthesizer with an American English voice (Dutoit et al., 1996). Half of the 256 possible words conformed to the "HH (height-height) pattern": V_1 was high if and only if V_2 was high. A different half, overlapping the first, conformed to the "HV (height-voice)" pattern: V_1 was high if and only if C_2 was voiced. Nine of the 18 American English native speakers who participated were assigned to the HH group, and nine to the HV group. The experiment consisted of a training phase and a test phase. In the training phase, the participant listened four times to a list of 32 pattern-conforming words (randomly chosen for each individual participant) and repeated each one aloud into a microphone. Exactly half of the 32 words conformed to the other pattern. In the test phase, the participant heard 32 pairs of new words, one pattern-conforming and one not, and chose the one most likely to be "a word of the language you studied".

Since the goal was to compare *learning* between the two groups, it was important that pre-existing (e.g., English-based) preferences for words conforming to one pattern not be mistaken for better learning of that pattern. To this end, the same test pairs were used for both groups. Half of the test pairs pitted an HH-conforming word against an HV-conforming one, while the other half had an HH- and HV-conforming word against a word that conformed to neither pattern. This design made it possible to measure pre-existing preference for the HH pattern using data from the participants in the HV group (who received equal numbers of HH-conforming and HH-nonconforming training stimuli), and vice versa. Perceptual accuracy was checked by blind-transcribing half of each participant's spoken repetitions and comparing them to the intended stimulus. The relevant features matched well over 90% of the time. Participants who had studied a language with a relevant dependency (e.g., height harmony) were replaced, as were those whose answers to a written post-experiment questionnaire showed that they had explicitly detected the pattern.

In subsequent experiments, the HV condition remained the same as in Experiment 1, and other dependencies were substituted for HH. Experiment 2 used voice-voice (VV); Experiment 3 used height of V_1 and backness of V_2 (HB); and Experiment 4 used place of C_1 and voicing of C_2 (PV). The experimental conditions are summarized in Table 2.

The results were analyzed using mixed-effects logistic regression with Participant as a random effect. The independent variables were chosen as follows: Each of the experiments in this series was

Experiment	V_1 height–C_2 voice vs. ...	Outcome
1	V_1 height–V_2 height	HH $>>$ HV ≈ 0
2	C_1 voice–C_2 voice	VV $>>$ HV ≈ 0
3	V_1 height–V_2 backness	HB ?$>$? HV ≈ 0
4	C_1 place–C_2 voice	PV ?$>$? HV ≈ 0

Table 1: Summary of Experiments 1–4.

modelled using a larger set of terms. The models were then reduced by backwards elimination. Any term which could not be eliminated from at least *one* of the models was retained (mutatis mutandis) in the analysis of all of them.

In Experiments 1–4, the HV group never chose the pattern-conforming test item with significantly more than chance frequency when the other factors were controlled. Participants trained on the HH or VV patterns in Experiments 1 and 2 did much better, by 0.716 and 0.736 logit units respectively (i.e., their odds of choosing the pattern-conforming item was more than twice as great). The statistical analysis showed that the effect was not due to rhyme or alliteration, and a subsequent experiment with a C_1-to-V_2 voice-height dependency showed that the superiority of HH and VV over HV was not explained by the greater salience of word-initial and word-final positions. Participants trained on the HB and PV patterns in Experiments 3 and 4 performed numerically, but non-significantly, better than the HV baseline (by 0.495 and 0.484 logits), and even this weak advantage was only found early in the test phase.

These results, summarized in Table 2, indicate an analytic bias: X–X dependencies enjoy a learning advantage over X–Y dependencies, regardless of what X and Y actually are. What properties must a learner have in order to behave this way?

3. The BLUMPS learner

The Bayesian Learner with Unbiased Multinomial Pattern Schemas (BLUMPS) was designed as a simple, tractable representative of the general class of phonotactic learners in which the final grammar is selected by first choosing a parametrized schema, then setting its parameters. A sophisticated example of a parametric-schematic learner would be one in which the competing schemas are sets of Optimality-Theoretic constraints, and the parameters of each schema are the ranking positions of the constraints.

BLUMPS is a "pure phonotactic learner" in the sense of Hayes (2004); its goal is to learn to distinguish well- from ill-formed surface strings. The competitors in BLUMPS are called *multinomial pattern schemas*. A pattern schema is like a bag of (possibly loaded) polyhedral dice, each responsible for a different part of the stimulus. Any given stimulus is thus associated with a probability that it will be generated when the dice in that bag are rolled. The schema is parametrized by how the dice are loaded. The learner is intrinsically unbiased between bags, and, within a bag, between different ways to load the dice. Its goal (roughly speaking) is first to choose the bag which makes the training data most likely when averaged over all possible loadings, then to find the best loading.

In some schemas, two featurally-congruent parts of the stimulus are accounted for by two different dice with the same faces, which may be loaded differently. In others, one die is rolled twice to generate the two parts; this is *parameter sharing*. The single die is BLUMPS's analogue of a grammar module; e.g., a schema might have a single die responsible for all and only consonants.

BLUMPS is designed to exploit the "Bayesian Occam's Razor" (MacKay, 2003, 343ff.), the principle that hypotheses with fewer adjustable parameters can fit fewer training sets, but assign higher probability to those which they do fit. The closest relative of BLUMPS in the linguistics literature, so far as I know, is the Minimum Description Length phonotactic learner of Ellison (1994), which heuristically penalizes analyses with more parameters. A concise review of Bayesian methods in linguistics can be found in Goldwater (2007). The rest of this section of the paper describes the components of BLUMPS in detail.

3.1. Representations

Stimuli are represented as feature values on a fixed-width "retina", as shown in Figure 1. Each experimental stimulus is represented with 16 feature values, 4 for each segment. The features occur in

Feature	Retinal segment				Consonants				Vowels			
	C_1	V_1	C_2	V_2	k	t	g	d	æ	ɔ	i	u
voiced	±		±		−	−	+	+				
aspirated	±		±		−	−	+	+				
Coronal	±		±		−	+	−	+				
Dorsal	±		±		+	−	+	−				
high		±		±					−	−	+	+
low		±		±					+	+	−	−
back		±		±					−	+	−	+
rounded		±		±					−	+	−	+

Figure 1: The representational "retina" and featural representation of segments.

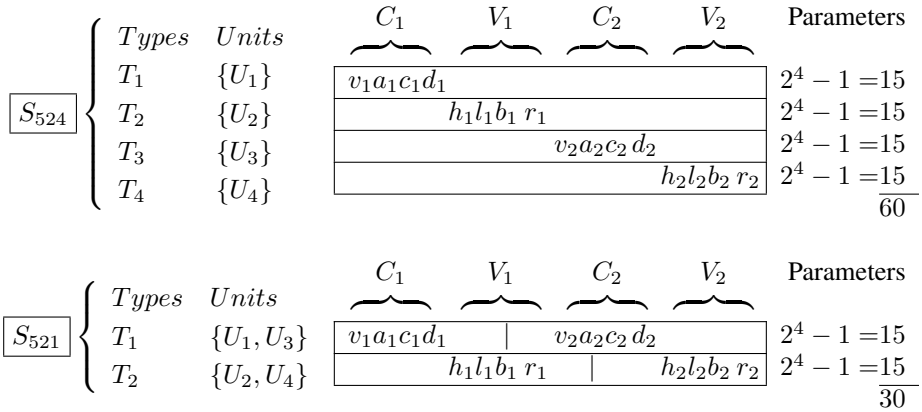

Figure 2: Two pattern schemas in which each segment is an independent unit. Parameters in S_{524} are unshared, while S_{521} has two instances of parameter-sharing.

redundant pairs, e.g., voicing and aspiration always have opposite values in the training data.[2]

3.2. Pattern schemas

Each competing explanation takes the form of a *pattern schema*, which is intended as a computationally tractable analogue of a parametrized phonotactic grammar. A pattern schema S begins with a partition of the retina into disjoint *units* $U_1, \ldots, U_n$—the "dice" of Section 3. Figure 2 shows S_{524}, which partitions the retina into four units, each containing all of the retinal positions belonging to a single segment.[3]

Since each feature can be either 0 (–) or 1 (+), a unit with k features ranges over 2^k different values, corresponding to binary numbers from 0 to $2^k - 1$. Given a stimulus s, the schema parses it into one value for each unit: $(u_1, \ldots, u_n)$. The phonotactic probability assigned to s by the schema S is just the product of the probabilities assigned by the units: $\Pr(s \mid S) = \prod_j Pr(u_j \mid U_j)$; i.e., the schema asserts that the units are statistically independent of each other. Pattern schemas are thus a kind of product partition model (Hartigan, 1990; Dahl, 2003).

For any k-feature unit, we can define any statistical dependency (gradient or absolute) among its features by specifying the $2^k - 1$ parameters $\mathbf{p} = (p_1, \ldots, p_{2^k-1})$, where $p_r = \Pr(u = r \mid U)$—the loading of the die.[4] Each setting of the parameters defines a multinomial distribution over the possible values of U. If the unit has been trained on a data set D^{old}, and is now asked to estimate the probability of a different set D^{new}, that estimate is obtained by finding the probability of D^{new}

[2]The redundancy is crucial; if only 8 features are used, the savings due to parameter-sharing is not enough make BLUMPS perform differently in the height-height and height-voice conditions of Experiment 1.

[3]The schemas are indexed according to the order in which they are generated by a particular algorithm, a deterministic variant of Algorithm 2.

[4]The reason there are $2^k - 1$ parameters rather than 2^k is that, since the probabilities have to add to 1, $\Pr(u = 00\ldots00 \mid U) = 1 - \sum_r \Pr(u = r \mid U)$.

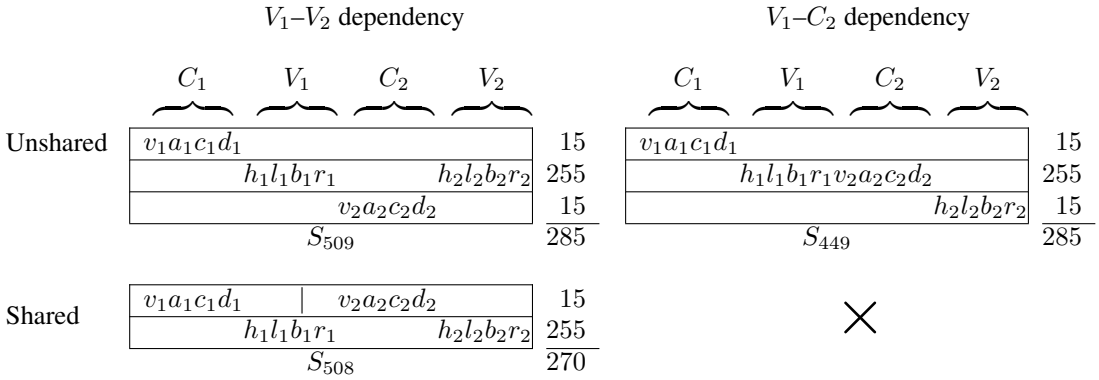

Figure 3: Importance of the residue. *Left*: A V_1–V_2 dependency leaves a symmetrical residue, which supports parameter-sharing. *Right*: A V_1–C_2 dependency leaves an asymmetric residue, and no sharing is possible.

for each value of $\mathbf{p}$, weighting it by the probability of that particular value of $\mathbf{p}$, and adding them all up: $\Pr(D^{new} \mid U, D^{old}) = \int_{\mathbf{p}} \Pr(D^{new} \mid U, \mathbf{p}) \Pr(\mathbf{p} \mid U, D^{old}) d\mathbf{p})$. Since $\Pr(D^{new} \mid U, \mathbf{p})$ is a multinomial distribution, $\Pr(\mathbf{p} \mid U, D^{old})$ is a Dirichlet distribution, and

$$\Pr(D^{new} \mid U, D^{old}) = \frac{\Gamma(\sum_r d_r^{new})}{\Gamma(\sum_r d_r^{new} + d_r^{old})} \prod_r \frac{\Gamma(d_r^{new} + d_r^{old})}{\Gamma(d_r^{old})} \tag{1}$$

where d_r^{old} and d_r^{new} are the number of occurrences of the rth value of u in D^{old} and D^{new}, and $\Gamma(n) = (n-1)!$ (Minka, 2003). In the training phase, D^{old} is empty and D^{new} is the training data. In the test phase, D^{old} is the training data and D^{new} is the test data. The actual learning rule used by BLUMPS is

$$\Pr(D^{new} \mid U, D^{old}, \rho) = \frac{\Gamma(\sum_r \rho d_r^{new})}{\Gamma(\sum_r \rho d_r^{new} + \rho d_r^{old} + 1)} \prod_r \frac{\Gamma(\rho d_r^{new} + \rho d_r^{old} + 1)}{\Gamma(\rho d_r^{old} + 1)} \tag{2}$$

BLUMPS adds 1 to each d_r^{old} so that in the training phase, when D^{old} is empty, the learner assigns equal probability to all values of $\mathbf{p}$; i.e., BLUMPS imposes a symmetric prior. Both d_r^{old} and d_r^{new} are multiplied by a learning parameter ρ, which allows control of the learner's receptiveness to data; see Section 5.3 below.

3.3. Parameter sharing and the residue

Some partitions of the retina create units which are *featurally congruent*: They are the same size and have the same features in the same order. Since the units range over the same possible values, and their parameters have the same interpretation, we allow the option of *parameter-sharing*. This is done by assigning every unit in a schema to a *type*, with the proviso that two units can be in the same type only if they are featurally congruent.

Parameter-sharing implements the BLUMPS analogue of grammatical modules. In Figure 2, for example, S_{521} asserts that the two consonants independently conform to one set of rules (a sub-grammar for consonants), and the two vowels independently conform to a different set of rules (a sub-grammar for vowels).

In order to capture a dependency in the data, a schema has to put the dependent parts of the retina into the same unit. If the dependency links featurally-congruent parts of the retina, the residue (the part of the retina remaining when the dependent unit is subtracted) can also be partitioned into featurally-congruent units which can be modularized by parameter-sharing. Dependencies between featurally-*in*congruent parts of the retina leave an asymmetric residue which blocks parameter-sharing. An example is shown in Figure 3. The within-tier V_1–V_2 dependency allows the consonants to share parameters, while the cross-tier V_1–C_2 dependency does not.

3.4. Training the learner

BLUMPS searches for pattern schemas which explain the training data. Since the search space is enormous—there are more than 10^{10} ways to partition a 16-bit retina, even without parameter-sharing—it cannot be searched exhaustively. Instead, the learner uses an evolutionary algorithm (Eiben & Smith, 2003). The search is controlled by three parameters: the mutation rate μ, the scale factor ρ_{train}, and the stability timeout T. The initial population is a random sample drawn from the space of all possible hypotheses by repeatedly applying Algorithm 2, as described below in Section 5.1. The population, whose size is fixed, completely replaces itself every generation by asexual reproduction. Opportunities to reproduce are raffled off to the schemas in the current population. If $\Pr(S_i)$ represents the proportion of S_i in the current generation, then the probability that S_i will win any given raffle is determined by Bayes's Rule:

$$\Pr(S_i \mid D^{train}, \rho_{train}) = \frac{\Pr(D^{train} \mid S_i, \rho_{train}) \Pr(S_i)}{\sum_j \Pr(D^{train} \mid S_j, \rho_{train})} \tag{3}$$

Each birth is subject to mutation. A single mutation happens in two stages. First, a feature chosen with uniform probability is deleted from the schema and then reinserted into a unit (which may be a new empty unit) chosen with uniform probability. Then the types are adjusted: Types which now contain incongruent units are split, and the source and destination unit of the moved feature are deleted from their respective types and reinserted into randomly chosen congruent types (which may be new empty types). The number of mutations per birth is Poisson-distributed with parameter μN, where μ is the mutation rate per feature and N is the number of features on the retina. In the simulations discussed here, BLUMPS was used as a batch learner; i.e., every generation used all of the training data. The evolutionary process continues until T generations have passed without improvement in the fitness of the fittest schema in the population.

3.5. Testing the learner

The probability of choosing the positive (pattern-conforming) test item from a pair (d_j^+, d_j^-) based on a single schema S_i was assumed to follow from the Luce choice rule (Luce, 1959 [2005], Ch. 1).

$$\Pr(+ \mid d_j^+, d_j^-, S_i, D^{train}, \rho_{test}) = \frac{\Pr(d_j^+ \mid S_i, D^{train}, \rho_{test})}{\Pr(d_j^+ \mid S_i, D^{train}, \rho_{test}) + \Pr(d_j^- \mid S_i, D^{train}, \rho_{test})} \tag{4}$$

The probability that the learner would choose the positive test item was defined as the probability that each schema would do so, weighted by the frequency of that schema in the final population:

$$\Pr(+ \mid d_j^+, d_j^-, D^{train}, \rho_{train}, \rho_{test}) = \sum_i \Pr(S_i \mid D^{train}, \rho_{train}) \Pr(+ \mid d_j^+, d_j^-, S_i, D^{train}, \rho_{test}) \tag{5}$$

Two different scale factors are used: ρ_{train} in the training phase controls schema fitness, and ρ_{test} in the test phase controls how precisely the schemas in the final population are adjusted. (See below, Section 5.3).

4. Simulations

We are now ready to address the main questions. First, does the effect actually happen as predicted? Does BLUMPS acquire X–X dependencies better than X–Y dependencies, given equivalent training data? Second, is the effect big enough to matter? If trained and tested using the same materials as the human participants, does BLUMPS show human-sized bias? To answer these questions, BLUMPS was

	Humans			Simulation		
Coefficient	Estimate	SE	$Pr(>\mid z \mid)$	Estimate	SE	$Pr(>\mid z \mid)$
(Intercept)	0.2742	0.1961	0.1620	0.2622	0.1091	0.0288 *
Studied HH	0.7161	0.2788	0.0103 *	0.8315	0.1543	<0.0001 ***
$V_1 = V_2$	–0.2596	0.2054	0.2062			
2nd half	–0.2788	0.2417	0.2488			
Studied HH $\times$ *2nd half*	–0.0598	0.3539	0.8659			
HH-nonconforming	0.1015	0.1314	0.4400			
1st in pair	0.4650	0.1768	0.0085 **			

Table 2: Actual (human) and simulated (BLUMPS) performance in Experiment 1, height-voice vs. height-height.

	Humans			Simulation		
Coefficient	Estimate	SE	$Pr(>\mid z \mid)$	Estimate	SE	$Pr(>\mid z \mid)$
(Intercept)	–0.0992	0.1975	0.6153	0.1550	0.1327	0.260
Studied HB	0.4958	0.2846	0.0815 .	0.1350	0.1877	0.482
2nd half	0.1040	0.2392	0.6636			
Studied HB $\times$ *2nd half*	–0.5830	0.3432	0.0894 .			
HB-nonconforming	–0.1157	0.1196	0.3336			
1st in pair	0.4559	0.1718	0.0080 **			

Table 3: Actual (human) and simulated (BLUMPS) performance in Experiment 2, height-voice vs. height-backness.

applied to Experiments 1 and 3. This section of the paper presents the results of simulations that behaved as desired; *why* they did so is discussed in Section 5

BLUMPS was implemented in R (R Development Core Team, 2005). It was trained and tested using the same materials as each of the 18 participants, with the stimuli presented as character strings rather than audio. In both simulations, the population of schemas was initialized by starting with 500 copies of S_{5172} (a 12-parameter schema shown in Figure 6) and mutating each one 100 times to yield a random sample from the space of possible schemas (see below, Section 5.1). In the training phase, the learner read all of the training words before any learning took place; i.e., it was run in batch mode. Once the training data had been read, the evolutionary algorithm repeatedly selected, bred, and mutated the schemas, until T = 100 generations had passed without improvement in the fitness of the fittest schema. The mutation rate was μ = 2/16 mutations per feature, for an average of two mutations per birth.

All of the foregoing are uninteresting "nusiance parameters"; they have to be set properly so that BLUMPS can find the fittest schema before the simulation is stopped, but do not otherwise affect performance. The scale parameters ρ_{train} and ρ_{test}—the only free non-nuisance parameters—were set at 0.175 and 0.5 respectively, values which were found by trial and error to give a good match to the human data.

Simulation results for Experiment 1 are shown in Table 4, side by side with those of the human participants. The parameter estimates are expressed in terms of the logarithm of the odds ratio of a correct (pattern-conforming) response. Several independent variables were relevant to humans but not to BLUMPS, since BLUMPS is incapable of learning or forgetting during the test phase, does not tend to prefer the first stimulus of a test pair, and has no native-language experience to bias it in favor of particular patterns. The remaining terms are the intercept, which is the log-odds of a correct response in the HV group, and *Studied HH*, which is the difference in log-odds between the HV and HH groups. As Table 4 shows, these are nearly the same for humans and BLUMPS. The significance levels are higher for BLUMPS, because the simulated subjects differ less amongst themselves than the humans do, and because the model has fewer terms. Table 4 compares human and simulated performance in Experiment 3. In both cases, the HV group does poorly, whereas the HB group is non-significantly better.

These results confirm that BLUMPS can match the human learning advantage for the height-height pattern over the height-voice one, and the lack of advantage for the height-backness pattern. Experiments 2 and 4 are merely Experiments 1 and 3 with the features permuted. Permuting the features does not affect BLUMPS's performance, and did not affect human performance (see Section 2), so BLUMPS is equally good at matching the human data in those experiments.

Algorithm 1: Mutating the underlying partition.

Input: A partition P of $\{1, \ldots, N\}$ into nonempty disjoint units, $S = \{U_1, \ldots, U_m\}$.
Output: A (possibly different) partition P' of $\{1, \ldots, N\}$ into nonempty disjoint units, $P' = \{U'_1, \ldots, U'_{m'}\}$.

1 $P^* \leftarrow P$;
2 $b \leftarrow$ random element of $\{1, \ldots, N\}$;
3 $source \leftarrow$ subscript of unit of P^* containing b;
4 delete b from U^*_{source};
5 **if** $U^*_{source} = \{\}$ **then**
6 delete U^*_{source} from P^*;
7 **end**
8 $J \leftarrow$ {subscripts of remaining units of P^*};
9 $dest \leftarrow$ random element of $J \cup \{0\}$;
10 **if** $dest = 0$ **then**
11 $P' \leftarrow P^* \cup \{b\}$;
12 **else**
13 $P' \leftarrow \{U^*_1, \ldots, U^*_{dest} \cup \{b\}, \ldots, U^*_{|P^*|}\}$ (excluding U^*_{source} if deleted);
14 **end**
/* Return b for the convenience of Algorithm 2. */
15 **return** P', b;

5. Analysis of the learner

The simulations show that BLUMPS can mimic human performance. This section of the paper investigates whether BLUMPS does that for the intended reason, namely, the superior explanatory power of schemas which use parameter-sharing.

5.1. Mutation algorithm

The mutation algorithm is important in two respects. First, it is used to randomize initial population by repeated mutation of multiple copies of a single schema. Second, it is responsible for innovation during the training phase. This section of the paper shows that the algorithm is unbiased with regard to the partition of the retina into units, and does not favor schemas with parameter-sharing over those without.

5.1.1. Effect on partition into units

In the long run, the mutation algorithm gives equal frequency to all possible partitions of the retina into units, regardless of the initial population of schemas. To see this, consider Algorithm 1, which is the half of the mutation algorithm which handles the units. Algorithm 1 defines transition probabilities between partitions, i.e., between schemas when the types are ignored. Any partition P_1 can be transformed into any other partition P_2 in $N - 1$ mutations,[5] so Algorithm 1 defines an ergodic Markov chain in the space of all possible partitions of $\{1, \ldots, N\}$. Hence, if Algorithm 1 is applied enough times, the long-term frequency of any given partition will be independent of the initial state.

What are those long-term frequencies? There is a transition from P to P' iff there is a b such that deleting b from P and from P' yields the same subpartition P^* at Step 8. When the algorithm subsequently re-inserts b, the result is equally likely to be P or P', so the transition matrix is symmetric: $\Pr(P \to P') = \Pr(P' \to P)$. The long-term frequencies are therefore identical for all Ps Randall (2006). Algorithm (**??**) is just a Markov Chain Monte Carlo procedure for sampling random set partitions, and, in the absence of selection, the BLUMPS mutation algorithm will tend to produce a uniform, unbiased distribution of partitions.

[5] Let $P = P_1$. For b in $\{2, \ldots, N\}$, let b^* be the smallest element of b's unit in P_2. In P, move b to the unit containing b^*, or to a new singleton unit if $b^* = b$. The final P is equal to P_2.

k	1	2	3	4	5	6	7	8	9–16	All
n_k	2.111	2.185	1.477	0.731	0.283	0.086	0.023	0.005	< 0.001	6.930

Table 4: Mean theoretical distribution of unit sizes at equilibrium, where n_k is the average number of k-feature units per schema.

How long is the "long term"? In a uniform distribution of partitions, the average number of units of size k per partition is given by $\hat{n}_k = \binom{N}{k} B_{N-k}/B_N$, where B_n is the nth Bell number (Knuth, 2005, 73). We can use the difference between the theoretical and actual frequencies as a measure of convergence. Let $\delta = (\sum_{k=1}^{N} \frac{1}{N}(\hat{n}_k - n_k)^2)^{\frac{1}{2}}$, where n_k is the empirical average. Figure 4 shows how δ develops for a population of 500 schemas on a 16-bit retina initialized to the one-big-unit schema (dashed curve), the 16-different-units schema (dotted curve), and S_{5172}, the 8-unit schema actually used as the initial state in these simulations (solid curve). It appears that in the absence of selection, the mutation algorithm effaces the initial state by about the 75th generation.

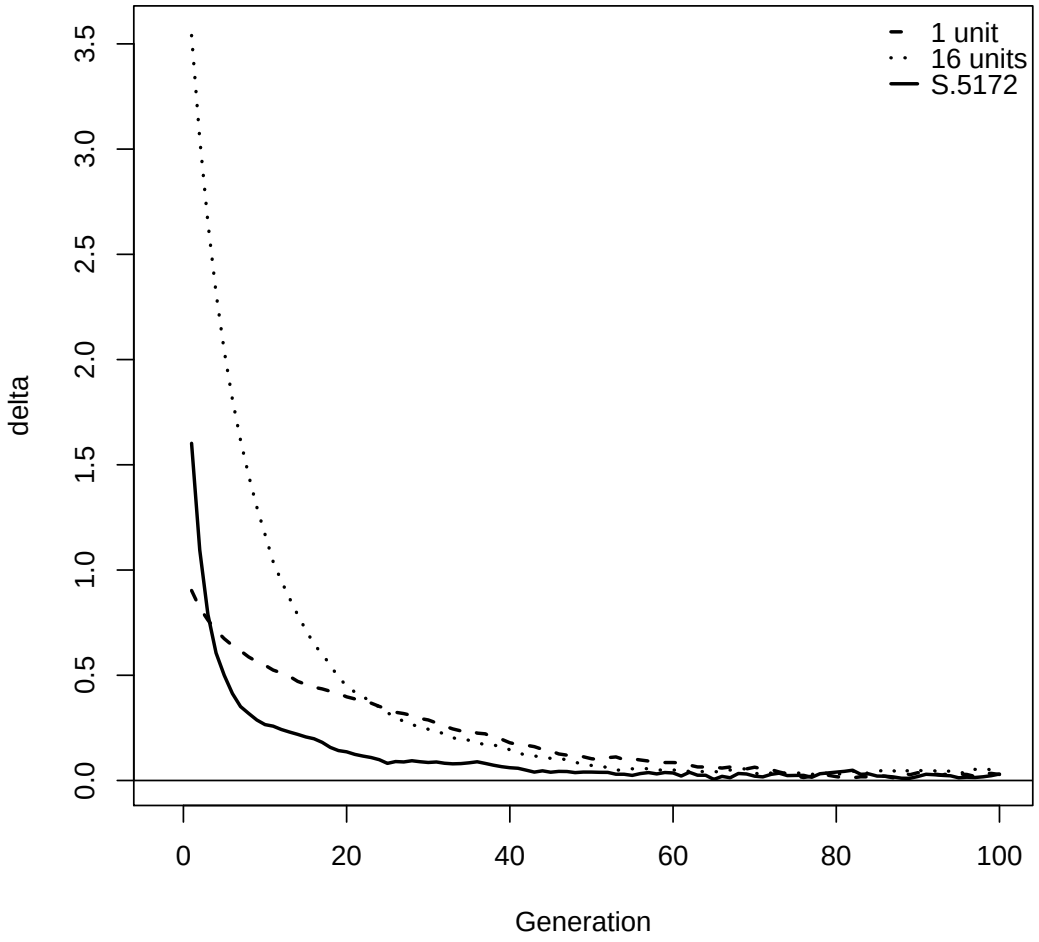

Figure 4: Convergence of Algorithm 1.

5.1.2. Effect on parameter-sharing

The next question is whether mutation, without selection, is biased for or against parameter-sharing. For any unit, there can be at most one other featurally-parallel unit, since every feature on this retina occurs exactly twice. Hence, whenever Algorithm 2 gets to Line 17, J has at most one element (and may have none, if the now-typeless source and destination units are congruent with each other). The probability of choosing "0" and putting U'_{source} or U'_{dest} into a new type by itself is therefore at least 1/2. Thus, an opportunity for parameter-sharing is less than 50% likely to be taken advantage of. Given the distribution of unit sizes in Table 4, opportunities will not be frequent, and when they do occur will mostly involve one- and two-feature units.

5.2. Fitness

Long-term departures from the mutation algorithm's equilibrium population can only be due to fitness-based selection. Figure 5 illustrates the dependence of fitness on training data. Only the 5187 schemas which do not put redundant features (e.g., voicing/aspiration) into separate units are shown; the others, of which there are more than 10^{10}, are too unfit to matter. For each schema, the axes show the logarithm of its average fitness when trained in each condition of Experiment 1 using ρ_{train} = 0.175.

The schemas cluster along three parallel lines. Those which can capture the height-height dependency but not the height-voice one have greater fitness in the HH condition than in the HV one, and form the upper cluster, while those for which the reverse is true form the lower one. Schemas which can capture both dependencies or neither are equally fit in both conditions and lie along the line $y = x$. Schemas with more parameters are plotted with smaller symbols, and those with parameter-sharing are plotted as squares rather than circles. It is clear that the fittest schemas are, as expected, those which economize on parameters by sharing—the large squares. The fittest schemas in the upper, middle, and lower clusters are S^{HH}_{5084}, S_{5172}, and S^{HV}_{4566}, shown in Figure 6. The outcome of the simulation is determined by their relationship to each other.

5.3. Training, convergence, and the final population

For a learner trained in the HH condition, S^{HH}_{5084} is the fittest schema of all, with the neutral S_{5172} coming in second. The fitness difference is larger than the log scale of Figure 5 makes it seem: 4.9×10^{-67} vs. 8.4×10^{-68}, or nearly six times greater. The final population is therefore dominated by S^{HH}_{5084}.

For a population of 500 schemas with $\rho_{train} = 0.175$ and $\mu = 2/16$, BLUMPS reaches its final state no later than the 25th generation, and stays there until the convergence timeout is reached. The final population contains, on average, only 67 copies of S^{HH}_{5084}, which get to do virtually all of the breeding, plus an entourage of much-less-fit mutants of S^{HH}_{5084}—about 135 with a single mutation, another 135 with two, and the remaining 161 with more (since mutation is Poisson-distributed with parameter μN = 2), The mutants are unfit because they are almost certain to leave at least one redundant feature pair, like v_1 and a_1, in two different units. However, even a very unfit mutant will still make the right choice in the test phase as long as at at least one of h_1 or l_1 still shares a unit with at least one of h_2 or l_2.

In the HV condition, S_{5172} is now the fittest schema, leading S^{HV}_{4566} by 1.7×10^{-67} to 3.2×10^{-68}, or a factor of more than five. Convergence is slower, with the final state being reached no later than the 125th generation. The neutral S_{5172} dominates the final population in the same way that S^{HH}_{5084} did in the HH condition. Neither it nor most of its mutant entourage can capture the height-voice dependency, so the learner makes wrong choices in the test phase.

The scale factor ρ_{train} determines how much influence the training data exerts on fitness. Data is presented to the learner in the form of frequency counts for each of the 2^{16} possible stimuli, multiplied by ρ_{train}. Reducing ρ_{train} in effect gives the learner data with the same proportions, but less of it. Very few schemas are consistent with hearing 200 height-harmonic stimuli in 200 trials, but nearly all are consistent with 2 out of 2. Hence, reducing ρ_{train} improves and equalizes all schemas' ability to fit the training data. As ρ_{train} is increased from zero in Experiment 1, the learner goes through four different stages:

- $\rho_{train} \approx 0$: As all schemas have nearly equal fitness, the final population reflects the equilibrium

Algorithm 2: Mutating the whole schema.

Input: A partition S of $\{1, \ldots, N\}$ into nonempty disjoint units, $\{U_1, \ldots, U_m\}$, which are grouped into nonempty disjoint types $\{T_1, \ldots, T_n\}$, where two units belong to the same type only if they are featurally congruent.

Input: A possibly different partition S' of similar description.

/* Mutate the underlying partition. */

$P \leftarrow \{U_1, \ldots, U_m\}$;

$P', b \leftarrow$ result of running Algorithm 1 on P;

/* Identify source (possibly empty), destination, and bystander units. */

$U_{source} \leftarrow$ unit of P which contains b;

$U'_{source} \leftarrow U_{source} - \{b\}$;

$U'_{dest} \leftarrow$ unit of P' which contains b;

$bystanders \leftarrow P' - U'_{source} - U'_{dest}$;

/* Bystander units stay in their old types. */

$n' \leftarrow 0$;

for $i \in \{1, \ldots, n\}$) **do**

  $T' \leftarrow T_i \cap bystanders$;

  if $T' \neq \{\}$ **then**

    $n' \leftarrow n' + 1$;

    $T'_{n'} \leftarrow T'$;

  end

end

/* Changed units are randomly assigned to congruent types, or to a new type. */

foreach $U' \in \{U'_{source}, U'_{dest}\}$ **do**

  $J \leftarrow \{j \mid U'$ is congruent with $T'_j\}$;

  $j \leftarrow$ random element of $J \cup \{0\}$;

  if $j = 0$ *and* $U' \neq \{\}$ **then**

    $n' \leftarrow n' + 1$;

    $T'_{n'} \leftarrow \{U'\}$;

  else

    $T'_j \leftarrow T'_j \cup U'$;

  end

end

$S' \leftarrow \{T'_1, \ldots, T'_{n'}\}$;

return S';

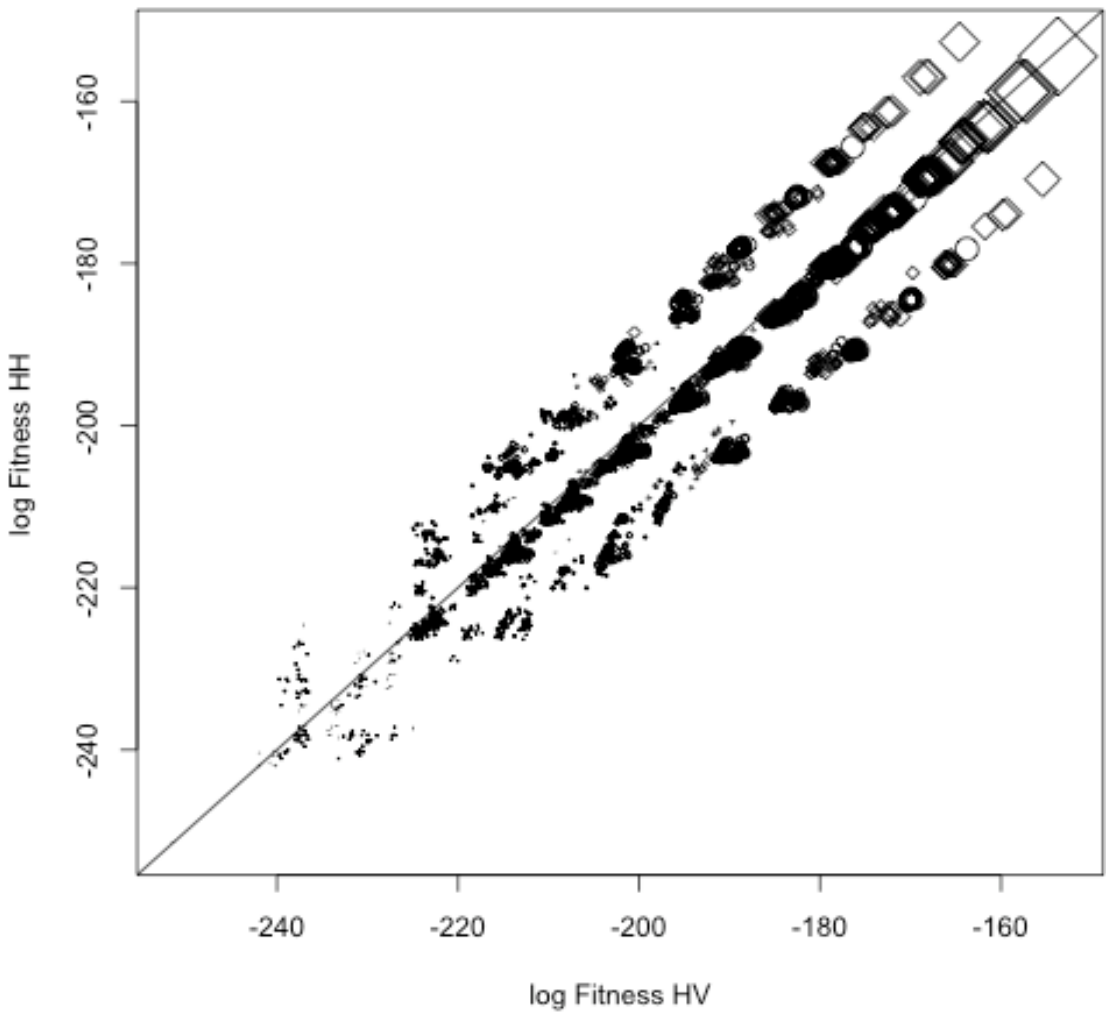

Figure 5: Log of average fitness of schemas in the height-voice vs. height-height conditions of Experiment 1 for learning parameter $\rho_{train} = 0.175$. Symbol diameter is inversely proportional to number of parameters. Schemas with shared parameters are plotted with squares, others with circles.

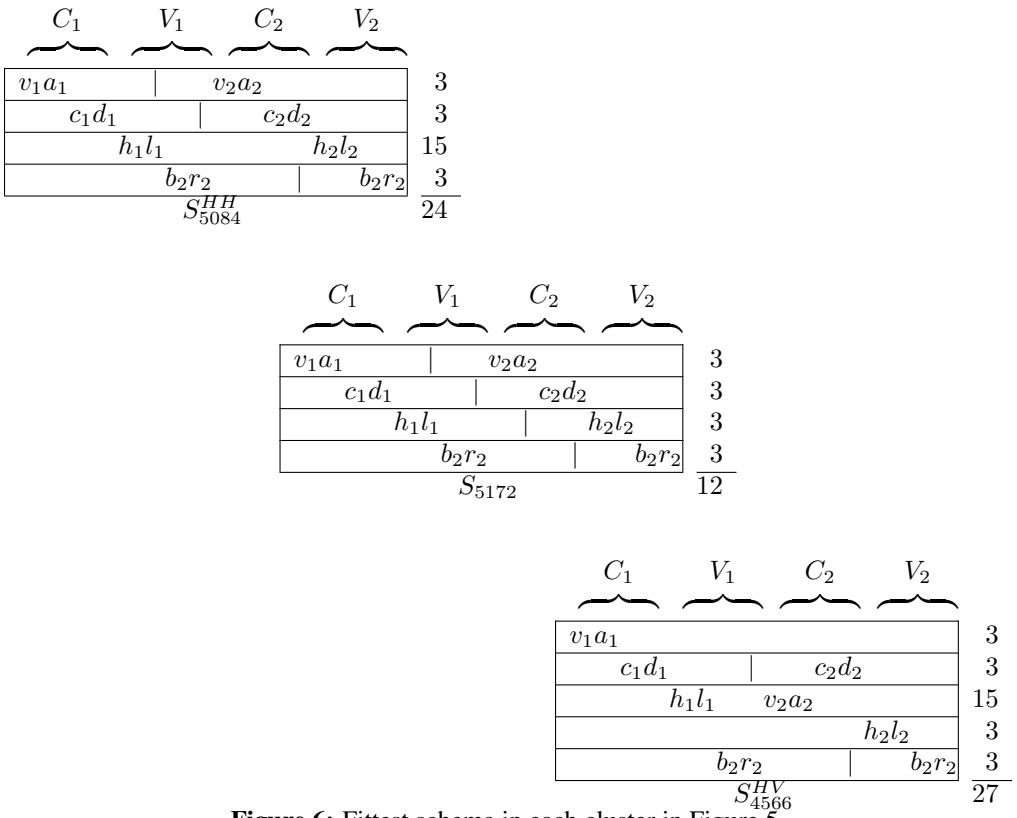

Figure 6: Fittest schema in each cluster in Figure 5.

	ρ_{test}				
Coefficient	0.1	0.25	0.5	0.75	1.00
(Intercept)	0.0404	0.1201	0.2622	0.4059	0.5508
Studied HH	0.1305	0.3858	0.8315	1.2847	1.7410

Table 5: Effect of varying ρ_{test} in Experiment 1; ρ_{train} = 0.175.

distribution of the mutation algorithm. It contains many schemas with units large enough to capture the dependencies. so performance is somewhat above chance in both conditions.

- $0.05 < \rho_{train} <\sim 0.125$: In both conditions, the neutral schema S_{5172} dominates the final population, then gives wrong answers in the test phase, leading to equal near-chance performance.
- $\sim 0.125 < \rho_{train} <\sim 0.225$: The final population is dominated in the height-height condition by S^{HH}_{5084}, which gives correct test answers, but in the height-voice condition by S_{5172}, which gives incorrect ones. Result: Performance is much better than chance in the height-height condition, and near chance in the height-voice condition.
- $\rho_{train} >\sim 0.225$: The final populations are dominated by S^{HH}_{5084} and S^{HV}_{4566}, respectively, both of which give correct test answers. Performance in both conditions is much better than chance.

Higher values of ρ_{train} favor schemas which fit the data better, regardless of the expense in terms of parameters, until ultimately the fittest schema is the one with $2^{16} - 1$ parameters, which can fit the training data perfectly.

The other scale parameter ρ_{test} controls the sensitivity of the schemas in the final population to both the training and test data, with lower values leading to worse (more chance-like) performance. Table 5 illustrates the effects of varying ρ_{test} on performance in Experiment 1.

6. Discussion

Together, the Bayesian Occam's Razor and the parametric parsimony of modular grammars offer an explanation for modularity bias. Unlike the proposals cited in Section 1, this explanation does not appeal to pre-existing preferences between objects meeting different formal descriptions. BLUMPS does not say

(I)
"Favor height harmony over height-voice and height-backness interactions."

nor does it say

(II)
"Favor one-feature dependencies over two-feature dependencies."

nor even

(III)
"Favor schemas which have fewer adjustable parameters."

All it says is,

(IV)
"Favor schemas which make the training data probable",
i.e.,
"Favor good explanations."

This last is the only hard-wired learning preference that distinguishes height-height from height-voice and height-backness dependencies in BLUMPS. The simulations show that if (IV) is granted, (III), (II), and (I) follow, provided that X–Y dependencies require more parameters than X–X ones. This proviso holds in BLUMPS because of parameter-sharing: X–X dependencies leave a symmetrical residue which invites parameter-sharing, whereas X–Y ones leave an asymmetrical residue which blocks it. Parameter-sharing is BLUMPS's analogue of grammatical "modules", or subsystems. The advantage of X–X dependencies is that they mind their own business and do not disrupt the learner's analysis of whatever else is going on.

An interesting prediction follows: When the residue itself does not allow parameter-sharing, the X–X advantage should disappear. In Experiment 1, for example, the voicing and aspiration features of C_1 and C_2 had the same distribution in the training data, making schemas like S^{HH}_{5084} a good explanation since one set of parameters controls both pairs of features. If the distributions had differed from each other, S^{HH}_{5084} and similar schemas would have been much less fit, and the HH–HV difference would have been reduced.

This should not be understood to say that human learners have *no* phonetically- or phonologically-detailed analytic biases; there is laboratory evidence that they do (Schane et al., 1974; Wilson, 2003b; Carpenter, 2005; Wilson, 2006; Finley & Badecker; Chambers et al., 2008). The point is rather that modularity bias does not require them. A believable model of phonological learning will of course have to include all human analytic biases. One possible approach would be to add substantive biases to BLUMPS, either by modifying the fitness function to favor particular schemas, or by assigning non-symmetric prior distributions $\Pr \mathbf{p}$ over the parameters of the schemas, so as to, e.g., give vowel harmony an advantage over vowel disharmony.

On the other hand, the pattern schemas of BLUMPS are a construct of convenience, chosen to simplify the calculations; the crucial ideas of the Bayesian Occam's Razor and modular parametric parsimony are in principle applicable in many frameworks. For example, consider an Optimality-Theoretic learner in which the schemas are sets of constraints and the parameters are the constraint rankings. Suppose $\mathbf{C}$ is one such schema, a set of $m + n$ OT constraints, of which m apply only to segments, and n apply only to tones. It does not matter how the segmental constraints are ranked relative to the tonal constraints, only how the segmental and tonal constraints are ranked relative to each other;

hence, even though there are $(m+n)!$ ways to rank $\mathbf{C}$, there are really at most $m!+n!$ distinct grammars. If $\mathbf{C}'$ has the same number of constraints, but some of them link tones with segments, then it matters how segmental constraints are ranked relative to tonal constraints, and the number of possible distinct grammars is greater than that for $\mathbf{C}$. Hence, equivalent learning data may favor $\mathbf{C}$ more than $\mathbf{C}'$ Moreton (to appear). The same argument would apply to other learners in which constraints are induced from the data, and competition is between constraint *sets*—a situation different from that of current constraint-induction models, in which constraints are evaluated one at a time (Boersma & Pater, 2007; Hayes & Wilson, 2007).

A separate but related question is whether analytic modularity bias, as observed in the lab, contributes to the prevalence of modular patterns in natural-language phonology. The question is hard because channel biases, such as height coarticulation, also tend to relate similar elements, so that many natural-language modular patterns could owe their high frequency to either factor. The two possibilities may be distinguishable through the study of cases of "underphonologization"; for discussion, see Moreton (2008); Yu (to appear); Blevins (to appear).

References

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

Beddor, Patrice Speeter, Rena Arens Krakow & Stephanie Lindemann (2001). Patterns of perceptual compensation and their phonological consequences. Hume, Elizabeth & Keith Johnson (eds.), *The role of speech perception in phonology*, Academic Press, San Diego, chap. 3, pp. 55–78.

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

Blevins, Juliette (2006). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics* 32:2, pp. 117–165.

Blevins, Juliette (to appear). Phonetically-based sound patterns: typological tendencies or phonological universals? *Papers in Laboratory Phonology 10*, Mouton de Gruyter.

Boersma, Paul & Joe Pater (2007). Constructing constraints from language data: the case of Canadian English diphthongs. Handout, NELS 38, University of Ottawa.

Carpenter, Angela C. (2005). Acquisition of a natural vs. an unnatural stress system. Brugos, Allejna, Manuella R. Clark-Cotton & Seungwan Han (eds.), *Papers from the 29th Boston University Conference on Language Development (BUCLD 29)*, Cascadilla Press, Somerville, pp. 134–143.

Chambers, Kyle E., Kristine H. Onishi & Cynthia Fisher (2008). A vowel is a vowel: generalizing newly-learned phonotactic constraints to new contexts. MS, Reed College.

Clements, G. N. (1995). The geometry of phonological features. Goldsmith, John A. (ed.), *Phonological theory: the essential readings*, Blackwell, Malden, pp. 201–223.

Clements, G. N & Elizabeth V. Hume (1995). The internal organization of speech sounds. Goldsmith, John A. (ed.), *The handbook of phonological theory*, Blackwell, Boston, chap. 7, pp. 245–306.

Dahl, David B. (2003). Modal clustering in a univariate class of product partition models. MS, Department of Statistics, University of Wisconsin, Madison.

de Lacy, Paul (2006). Transmissibility and the role of the phonological component. *Theoretical Linguistics* 32:2, pp. 185–196.

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

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

Ellison, T. Mark (1994). The iterative learning of phonological constraints. *Computational Linguistics* 20:3.

Finley, Sara & William Badecker (). Right-to-left biases for vowel harmony: evidence from artificial grammar.

Frisch, Stefan, Janet B. Pierrehumbert & Michael B. Broe (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:1, pp. 179–228, URL `http://www.ai.mit.edu/projects/dm/featgeom/frisch-etal-similarity.pdf`.

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

Goldwater, Sharon J. (2007). *Nonparametric Bayesian models of lexical acquisition*. Ph.D. thesis, Brown University, Providence, Rhode Island.

Gordon, Matthew (2004). Syllable weight. Hayes, Bruce, Robert Kirchner & Donca Steriade (eds.), *Phonetically-based phonology*, Cambridge University Press, Cambridge, England, pp. 277–312.

Hansson, Gunnar Ólafur (2008). Diachronic explanations of sound patterns. *Language and Linguistics Compass* .

Hartigan, J. A. (1990). Partition models. *Communications in Statistics: Theory and Methods* 19:8, pp. 2745–2756.

Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. Kager, René, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, Cambridge University Press, Cambridge, England, chap. 5, pp. 158–203.

Hayes, Bruce & Colin Wilson (2007). A maximum entropy model of phonotactics and phonotactic learning. MS, UCLA. To appear in *Linguistic Inquiry*.

Kavitskaya, Darya (2002). *Compensatory lengthening: phonetics, phonology, diachrony*. Routledge, New York.

Kiparsky, Paul (2006). The amphichronic program vs. Evolutionary Phonology. *Theoretical Linguistics* 32:2, pp. 217–236.

Knuth, Donald E. (2005). *The art of computer programming*, vol. 4, Fascicle 3: Generating all combinations and permutations. Addison-Wesley, Upper Saddle River, New Jersey.

Luce, R. Duncan (1959 [2005]). *Individual choice behavior: a theoretical analysis*. Dover, Mineola, New York, reprint of original 1959 edition edn.

MacKay, David J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, England.

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

Minka, Thomas P. (2003). Estimating a dirichlet distribution. Technical report, Microsoft research.

Moreton, Elliott (2008). Analytic bias and phonological typology. *Phonology* 25:1.

Moreton, Elliott (to appear). Underphonologization and modularity bias. Parker, Steve (ed.), *Phonological argumentation: essays on evidence and motivation*, Equinox, London.

Newport, Elissa & Richard N. Aslin (2004). Learning at a distance i: statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48, pp. 127–162.

Ohala, John J. (1994). Towards a universal, phonetically-based theory of vowel harmony. *Proceedings of the International Conference on Spoken Language Processing*, pp. 491–494.

Onnis, Luca, Korin Richmond & Nick Chater (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language* 53, pp. 225–237.

Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted (2003). Phonological rule-learing and its implications for a theory of vowel harmony. Tsujimura, M. & G. Garding (eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)*, pp. 101–114.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Randall, Dana (2006). Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering* pp. 30–41.

Rose, Sharon & Rachel Walker (2004). A typology of consonant agreement as correspondence. *Language* 80:3, pp. 475–531.

Schane, Sanford A.., Bernard Tranel & Harlan Lane (1974). On the psychological reality of a natural rule of syllable structure. *Cognition* 3:4, pp. 351–358.

Wilson, Colin (2003a). Analytic bias in artificial phonology learning: consonant harmony vs. random alternation. Handout from presentation at the Workshop on Markedness and the Lexicon, Massachusetts Institute of Technology.

Wilson, Colin (2003b). Experimental investigation of phonological naturalness. Garding, G. & M. Tsujimura (eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)*, Cascadilla Press, Somerville, pp. 533–546.

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

Yip, Moira (2002). *Tone*. Cambridge University Press, Cambridge.

Yu, Alan C. L. (to appear). Tonal effects on perceived vowel duration. MS, University of Chicago. To appear in: Papers in Laboratory Phonology 10.