Inter- and intra-dimensional dependencies in implicit phonotactic learning

Elliott Moreton[a]

[a] *Department of Linguistics, University of North Carolina, Chapel Hill, NC 27599–3155, U.S.A. Tel. 1-919-962-1192, fax 1-919-962-5839.*

Abstract

Is phonological learning subject to the same inductive biases as learning in other domains? Previous studies of non-linguistic learning found that intra-dimensional dependencies (between two instances of the same feature) were learned more easily than inter-dimensional ones. This study compares implicit learning of intra- and inter-dimensional phonotactic dependencies. A series of six unsupervised implicit-learning experiments shows that a pattern based on agreement between two instances of the same feature is easier to learn than one based on correlation between instances of two different features. The results are interpreted as evidence for domain-general restrictions on the form of domain-specific learning primitives.

Keywords: phonotactic learning, concept learning, implicit learning, inductive bias, complexity

1. Introduction

A major question in cognitive science is the existence or otherwise of cognitive specializations for language acquisition (Hume and Johnson, 2001; Saffran, 2003; Jackendoff and Pinker, 2005; Christiansen and Chater, 2008; Evans and Levinson, 2009). Since all learning relies on inductive biases that render some generalizations more salient to the learner than others (Pinker, 1979; Mitchell, 1990; Gallistel et al., 1991; Marler, 1991), one way to test the domain-specificity of learning processes is to compare inductive biases across learning domains.

The present study addresses the question of whether phonotactic learning and non-linguistic category learning share an inductive bias that renders within-stimulus dependencies more salient when they relate two instances of a single feature (e.g., nasality to nasality, or color to color) than when they relate instances of two different features (e.g., nasality to place of articulation, or color to shape). The difference between intra- and inter-dimensional dependencies is of particular theoretical significance because it plays a major role in phonological theory but almost none in general psychological models of concept learning. The situation in the laboratory is the other way around: There is considerably more evidence for an intra-dimensional advantage in non-linguistic concept learning than there is in phonological learning. This study aims to redress that imbalance.

1.1. Phonotactic learning as concept learning

A language learner acquires implicit knowledge of phonotactic generalizations about how phonemes combine to form larger units such as syllables and words. Phonotactic patterns can also be acquired in the lab: Participants familiarized with stimuli conforming to a particular pattern come to distinguish in performance between novel pattern-conforming and pattern-nonconforming stimuli. Such effects have been observed in learners as young as four months (Chambers et al., 2003; Saffran and Thiessen, 2003; Seidl and Buckley, 2005; Cristià et al., 2011), and in paradigms as diverse as phoneme restoration (Ohala and Feder, 1994), explicit categorization (Pycha et al., 2003; Wilson, 2003; Endress et al., 2005), allomorph selection (Peperkamp et al., 2006), speeded repetition (Onishi et al., 2002), induced speech errors (Dell et al., 2000; Goldrick, 2004; Warker and Dell, 2006), language-game responses (Wilson, 2006), and immediate recall (Majerus et al., 2004). These experiments are essentially concept-formation tasks in which participants learn to categorize stimuli, explicitly or implicitly, according to whether they conform to the target phonotactic pattern.

To what extent does phonotactic learning share inductive biases with the learning of non-linguistic concepts? There are two principal strands of inquiry, reviewed at length in Moreton and Pater (to appear). One line of work has demonstrated that phonotactic learning shares certain *formal* biases with non-linguistic concept learning. A phonological class that is defined by a single-feature affirmation (e.g., "initial consonant is voiceless") is easier to learn in the laboratory than a two-feature biconditional (e.g., "initial consonant is voiceless if and only if it is labial") (Saffran and Thiessen, 2003; Cristià and Seidl, 2008), which in turn is easier than a three-feature nested biconditional (e.g., "if the vowel is high, the initial consonant is voiceless if and only if it is labial, but if the vowel is not high, the initial consonant is voiceless if and only if it is not labial") (Pycha et al., 2003; Kuo, 2009).[1] This same order of difficulty has repeatedly been demonstrated in non-linguistic concept learning (Shepard et al., 1961; Neisser and Weene, 1962; Nosofsky et al., 1994; Feldman, 2000; Love, 2002; Smith et al., 2004). These biases, however, are so fundamental that the sharing of them does not go far towards resolving the question of shared mechanisms. No learning device is likely to predict any other difficulty order unless it is deliberately engineered to do so. We must instead look at biases that could reasonably be otherwise.

Another approach, therefore, has investigated the existence of *substantive* biases, i.e., those concerned with inductive problems peculiar to phonology that are unlikely to be part of the equipment of a domain-general learner, such as syllable structure (Schane et al., 1974), palatalization (Wilson, 2006), and stress assignment (Carpenter, 2010). The evidential situation in this area is far from clear; findings frequently conflict, even within the same study. Much of the uncertainty stems from unavoidable confounds between the abstract phonological patterns and the concrete phonetic features. It is difficult to tell whether a given effect is due to, e.g., phonological stress, or to the duration, amplitude, etc., which instantiate it.

The present study pursues a middle way, investigating inductive biases which

[1] These are only a few of the most directly relevant studies of complexity in artificial-phonology learning. Many more are reviewed at greater length in Moreton and Pater (to appear).

are neither so fundamental as to seem inescapable, nor so specific as to be inseparable from their instantiating features. In particular, we focus on within-stimulus dependencies of two kinds: "intra-dimensional" dependencies between two instances of a single feature, and "inter-dimensional" ones between instances of two different features.

1.2. Intra- and inter-dimensional dependencies in phonotactic learning

Conspicuously often, phonotactic patterns in the world's languages involve within-word dependencies between instances of the same phonological feature. For example, the place-of-articulation feature has the same value in any two adjacent consonants in a Japanese word (Vance, 1987, Ch. 5), and tends to have different values in any two successive consonants in an Arabic triliteral verb root (McCarthy, 1986; Frisch and Zawaydeh, 2001). Patterns of this sort are well known under such names as "agreement", "harmony", "assimilation", "dissimilation", etc (Baković, 2011; Archangeli and Pulleyblank, 2011; Rose and Walker, 2011). Complementary "disagreement" or "dissimilation" patterns are rarer but still widespread (Bye, 2011). Many models of phonological knowledge reflect this observation by giving a special status to intra-dimensional dependencies (Goldsmith, 1976; McCarthy, 1988; Rose and Walker, 2004; Alderete and Frisch, 2008). To the extent that these models are intended to describe human cognition, this special status amounts to a hypothesis that intra-dimensional dependencies are more salient to a learner than inter-dimensional ones.[2]

However, the natural-language facts alone do not prove that an intra-dimensional learning advantage exists. It could instead be that phonological patterns are derived from pre-existing phonetic patterns of coarticulation or of auditory confusability, and that those phonetic precursors tend to be intra-dimensional. For example, vowel height harmony may be derived from vowel-to-vowel coartic-

[2] This is not to say that inter-dimensional patterns are rare in natural language. Common inter-dimensional patterns include spirantization (Kirchner, 1998), assibilation (Kim, 2001), palatalization (Guion, 1996), and the lowering of tones by voiced obstruents ("depressor consonants", Moreton 2010), as well as the many phonetically-unsystematic "crazy rules" that are idiosyncratic to particular languages (Mielke, 2004).

ulation, in which the height of one vowel affects that of another because the two vowels are competing for the same physical resource (control of the tongue body). In one version of this alternative hypothesis, there is an inductive bias favoring phonological patterns which have phonetic precursors; in another, phonological patterns arise when phonetic precursors are misinterpreted as phonological by learners (for overviews of these positions, see Hayes et al., 2004; Hansson, 2008). The availability of these alternative explanations necessitates experiments that study learning more directly.

Laboratory studies of artificial-phonology learning have found some evidence that, although both intra- and inter-dimensional dependencies can be learned in the lab, intra-dimensional dependencies are more salient than inter-dimensional ones. Wilson (2003) familiarized English-speaking participants on trisyllabic stimuli in which the initial consonants of the second and third syllables were correlated, then collected two-alternative forced-choice decisions between stimuli which conformed to the pattern and foils which did not. Performance was better when the pattern was agreement or disagreement in nasality than when it related the nasality of one consonant to the place of articulation of the other, leading Pater and Tessier (2006) to suggest that the nasal-nasal pattern might be facilitated by its intradimensional nature. Using a similar paradigm, Moreton (2008), with English speakers, and Lin (2009), with Mandarin and Southern Min Chinese speakers, found better performance for height-height and voice-voice dependencies than for a height-voice dependency.

This evidence is not conclusive. The inference from the Wilson (2003) results depends on the assumption that the place of articulation of American English [l] is not dorsal, which is debatable on phonetic and phonological grounds (Sproat and Fujimura, 1993; Walsh Dickey, 1997). The Moreton (2008) and Lin (2009) studies are open to alternative interpretations in which the advantage is actually for vowel–vowel and consonant–consonant dependencies over vowel–consonant ones, or for dependencies involving salient word-initial and word-final segments over word-medial ones (Moreton and Pater, to appear). Finally, Kuo (2009), with Mandarin speakers, found no difference in learning performance between a place–place dependency and a place–aspiration one.

1.3. Intra- and inter-dimensional relations in non-phonological learning

Opportunities for intra-dimensional dependencies arise naturally in phonology because successive phonemes in an utterance inhabit the same low-dimensional feature space. This kind of stimulus space is rare outside of language and other communicative systems such as music or birdsong (Hockett, 1960).[3] Models of general non-linguistic concept learning reflect this ecological fact by making no distinction between intra- and inter-dimensional dependencies. All stimulus features are represented the same way, irrespective of any commonalities between their physical representations (Gluck and Bower, 1988; Anderson, 1991; Kruschke, 1992; Nosofsky et al., 1994; Love et al., 2004; Feldman, 2006). These models predict no advantage for intra- over inter-dimensional dependencies. Despite the lack of attention in the modelling literature, however, there is empirical evidence that non-phonological concepts are easier to learn when they are defined intra-dimensionally.

Intra-dimensional equality relations were studied by Ciborowski and Cole (1973) and Ciborowski and Price-Williams (1974). Stimuli were cards showing two adjacent figures. Each figure had one of three shapes (triangle, circle, or square) and one of three colors (red, white, or black). Participants were trained to sort the cards into two face-down piles. The target concept was determined by a rule that used one feature from each of the two figures. Intra-dimensional

[3] Utterances in a language concatenate discrete units which, at the phonological level of production and perception, instantiate particular values from the same small feature set. The set of possible within-stimulus dependencies is thereby restricted to correlations between features at particular loci. There are other stimulus spaces that share this property, such as tunes and poker hands, and they, like language, invite featural comparison within the stimulus ("three of a kind", "inside straight", "full house", etc.). However, many categories in the world are not produced by a discrete combinatorial schema. Objects are not always easily resolvable into discrete parts (e.g., ice cream), and the parts they do have may be perceived in terms of very different sets of features (e.g., the seeds, leaves, and bark that characterize a tree species).

A reviewer points out that visual perception may *impose* discrete combinatorial structure on an image by parsing it into shape elements drawn from a finite, featurally-defined repertoire ("geons", Biederman 1987). If this view is correct, we would expect easier learning for within-stimulus dependencies between two instances of the same shape feature than between instances of two different ones.

rules used the colors of the two figures, or their shapes (e.g., positive instances had two red shapes), while interdimensional rules used one color and one shape (e.g., positive instances had one red figure and one triangle). Rules could be either conjunctive (red *and* red; red *and* triangle) or disjunctive (red *or* red; red *or* triangle). Learning success, measured by the mean number of errors before criterion performance, was significantly better for intra-dimensional rules than for inter-dimensional ones. This result, originally obtained with college students in New York City, was replicated with schoolchildren in New York and in rural Hawaii, and (in some conditions) with Kpelle villagers in rural Liberia.

Intra-dimensional dependencies can also involve two different values of the same dimension. Rogers and Johnson (1973) used stimulus cards with two boxes on them. In the inter-dimensional condition, one box contained a color and the other a shape, and the target concept was red/triangle. There were two intra-dimensional conditions. In one, both boxes contained colors, and the target concept was red/yellow. In the other, both boxes contained shapes, and the target was triangle/circle. Six-year olds took about three times as many trials to reach criterion in the inter-dimensional condition as in the intra-dimensional ones. (Four-year-olds performed alike in both conditions.)

Inter-dimensionally, biconditionals (if-and-only-if, or exclusive-or) are routinely found to be at least as hard as disjunctions, and disjunctions to be harder than conjunctions, across a wide range of stimulus spaces and experimental procedures (Bruner et al. 1956, Ch. 6, Neisser and Weene 1962; Hunt and Kreuter 1962; Conant and Trabasso 1964; Haygood and Bourne 1965; King 1966; Snow and Rabinovitch 1969; Gottwald 1971a,b; Lee 1981; but see Bourne and O'Banion 1971). Ciborowski and Cole (1973), in the above-cited study, found that the conjunction-disjunction difference was abolished for intra-dimensional categories; e.g., "at least one red" was no harder than "both red", while "a red or a triangle" was harder than "a red and a triangle". Using stimuli that were strings of uniquely colored $+$ and $-$ signs, Laughlin and Jordan (1967) and Laughlin (1968) found that intra-dimensionally, biconditionals (e.g., "red $+$ if and only if blue $-$") were easier than disjunctions (e.g., "red $+$ or blue $-$") and no harder than conjunctions (e.g., “red + and blue −”) — a striking difference from the usual order found with inter-dimensional patterns..

Intra-dimensional biconditionals are a special case of the equality or inequality relations; e.g., “red + if and only if blue −” is equivalent to “red symbol $\neq$ blue symbol”. Equality was studied by Hunt and Hovland (1960), who used flag-like stimuli instantiating a six-dimensional space with four values on each dimension. The flag was divided into an upper and a lower half, so that three feature types had one instance in each half of the flag. Participants were trained using labelled positive and negative instances, which were chosen so that three intra-dimensional rules agreed on how to classify them: conjunction (e.g., “upper red stripe and lower black stripe”), disjunction (e.g., “crosses in the top row and/or fleurs-de-lis in the bottom row”), or equality (e.g., “same number of upper and lower figures”). Subsequent tests with new stimuli on which the three rules disagreed showed that participants used conjunction and equality about equally often, but seldom used disjunction.

1.4. Aims of the present study

The current state of our knowledge can be summarized as follows: For phonotactic learning, an intra-dimensional advantage has been proposed, but not demonstrated, whereas for general concept learning, it has been demonstrated, but not proposed (i.e., not incorporated into a learning model). The main goal of this study is to establish the existence or otherwise of an intra-dimensional advantage in phonotactic learning by replicating and extending the Moreton (2008) and Lin (2009) studies in order to systematically eliminate alternative hypotheses. Table 1 summarizes the differences between the experiments.

2. Experiment 1

Experiment 1 of Moreton (2008) found that participants were more likely to choose a novel pattern-conforming stimulus over a nonconforming foil when they had been familiarized on an intradimensional biconditional pattern than on an interdimensional one. Our Experiment 1 attempted to replicate this result.

Table 1: Familiarization conditions in Experiments 1–6.

Experiment	Condition	Intra-dimensional?	Within-tier?	Adjacent?
1	height-height	Y	Y (V)	N
	height-voice	N	N	Y
2	voice … height	N	N	N
	height-voice	N	N	Y
3	backness-backness	Y	Y (V)	N
	backness-voice	N	N	Y
4	height-backness	N	Y (V)	N
	height-voice	N	N	Y
5	voice-voice	Y	Y (C)	N
	height-voice	N	N	Y
6	place-voice	N	Y (C)	N
	height-voice	N	N	Y

Table 2: Consonants and vowels used in the $C_1V_1C_2V_2$ stimuli in Experiments 1–6.

Consonants C_1 and C_2

	Coronal	Dorsal
Voiced	d	g
Voiceless	t	k

Vowels V_1 and V_2

	Front	Back
High	i	u
Nonhigh	æ	ɔ

The stimulus space consisted of all 256 $C_1V_1C_2V_2$ strings for which $C_i \in$ [t d k g] and $V_i \in$ [i u æ ɔ]. Each of the four positions was described by a factorial combination of two binary features, place and voicing for the consonants, height and backness for the vowels (Table 2).

Each experimental pattern partitioned the stimulus space into two equal halves. The *height-height pattern* was an intradimensional biconditional, satisfied when the two vowels were both high or both nonhigh. The *height-voice pattern* was an interdimensional biconditional, satisfied when C_2 was voiced ([d g]) if and only if V_1 was high. Each participant was familiarized with stimuli conforming to one of the patterns. Both groups were then tested on their ability to choose a novel pattern-conforming stimulus over a novel non-conforming one in a two-alternative forced-choice task.

In order to distinguish learned from pre-existing preferences, each familiarization group was used as a control for the other. All of the familiarization stimuli in the height-height group were height-height-conforming, but half were height-voice-conforming and half height-voice-nonconforming, whereas the reverse was true in the height-voice group. The same test pairs were used for both familiarization groups. In half of the test pairs, one word was height-height, but not height-voice-, conforming and the other height-voice-, but not height-height-, conforming. In the other half, one word was height-height- and height-voice-conforming and the other was height-height- and height-voice-nonconforming. The untrained effects of height-height- and height-voice-conformity could thus be de-confounded from preferences acquired by familiarization.

Sensitivity to repeated phonemes within a stimulus could lead to higher performance in the height-height group than the height-voice group, since (owing to the small vowel inventory) stimuli with a repeated vowel make up half of the possible height-height-conforming stimuli. Participants could either enter the experiment with a pre-existing preference for repeated vowels, or could acquire such a preference from familiarization on the height-height pattern. The stimuli were counterbalanced to allow both of these effects to be tested for and modelled out. The complete design is shown along with the results in Table 3.

2.1. Method

2.1.1. Participants

Twenty-seven participants were recruited at the author's institution by means of posted flyers and mass email offering $7 for a half-hour experiment. Participants were required to be at least 18 years of age, native speakers of English, with no known speech or hearing disorders at the time of testing. Results from three participants were excluded from analysis (one correctly described the pattern afterwards, one turned in an incomplete debriefing questionnaire, and one was a backup participant whose data turned out to be unnecessary for the 24-participant target). The height-height group consisted of 6 female and 6 male participants with a mean age of 22.4 years (s.d. 4.6 years). The height-voice group had 6 female and 6 male participants with a mean age of 21.7 years (s.d. 4.1 years).

2.1.2. Stimuli

Stimuli were synthesized at a 16-kHz sampling rate and 16-bit resolution using the MBROLA concatenative diphone synthesizer's "US 3" male American English voice (Dutoit et al., 1996). The fundamental frequency was left at its default setting, a 123-Hz monotone. The nominal durations of the two consonants were set to 75 ms. In order to get V_1 and V_2 to have the same actual duration of about 160–170 ms, their nominal durations were set to 169 ms and 225 ms respectively. The diphone used to make the second-syllable [du] contained about 45 ms of aspiration after the burst. To prevent confusion with [tu], 26 ms of the aspiration was removed. A silent interval of nominal duration 100 ms was synthesized at the beginning of the stimulus, and another of 25 ms at the end; however, the word-initial diphones contained intrinsic initial silence as well. All stimuli were 674 ms long, except those ending in [du], which were 648 ms long. No amplitude normalization was done, in order to maintain the natural amplitude difference between high and low vowels.

For each of the 24 participants, a set of 32 distinct familiarization stimuli was randomly chosen such each pattern-conforming $V_1C_2V_2$ sequence occurred exactly once. This insured that exactly half of the height-height familiarization stimuli were height-voice-conforming, and vice versa. To generate a test set, 64 novel test items, were randomly chosen such that each logically possible $V_1C_2V_2$ sequence occurred exactly once. The test items were randomly arranged into 32 pairs, subject to the condition that half of the pairs match a height-height-conforming against a height-voice-conforming stimulus (e.g., [tiku] vs. [dætu], and half match a height-height- and height-voice-conforming stimulus against a height-height- and height-voice-nonconforming one (e.g., [tigu] vs. [dædu]). Twelve test sets were generated, each heard by one participant in each familiarization group.

2.1.3. Procedure

Participants were tested individually in a double-walled soundproof chamber (Ray Proof Corporation, Model AS-200). The participant was seated in front of an Apple Macintosh iBook laptop computer, which played the audio stimuli and collected the responses under the control of a program written in Java 2, Version 1.4.2_09 (Sun Microsystems). At the beginning of the experiment, the participant was told, orally and then again in writing, that he or she would first study words in an artificial language by hearing and repeating them, then be "tested on how well you can recognize them". The instructions were recapitulated in writing at the beginning of the familiarization phase and of the test phase.

The familiarization phase consisted of four blocks of 32 trials, with the same stimuli randomly rearranged in each block. At the start of a familiarization trial, a stimulus was played for the participant through binaural mono headphones (Altec Lansing). The participant repeated the stimulus into a microphone attached to the headphones. This response was digitized, recorded, and saved to the computer's hard disk. The participant ended the trial by using the mouse to click a button on the screen labelled "Next". The instructions were to match the pronunciation as closely as possible. No feedback was given, and there was no visual stimulus aside from the "Next" button.

The study phase was followed by a written reminder of the test-phase instructions, then by by the test phase, which consisted of 32 two-alternative forced-choice trials. On each trial, the two stimuli were played in random order, separated by 150 ms of silence, while the screen displayed two buttons labelled "1" and "2". The participant was instructed to decide which stimulus was more likely to be "in the language you studied", and to click the appropriate button. The mouse response was recorded, and the next trial began immediately.

The entire experiment lasted 10–12 minutes. As soon as it was over, the participant completed a written questionnaire. The first question asked whether he or she had noticed any patterns in how the words were formed in the artificial language, and, if so, what they were. The remaining questions asked for age, sex, language and linguistics background, and history of hearing or speech disorders.

2.2. Results and discussion

The results were analyzed using mixed-effects logistic regression with the *lme4* library of the *R* statistical package (R Development Core Team, 2005).[4] The dependent variable was the participant's response, coded as choice of the pattern-conforming or non-conforming test stimulus (1 or 0, respectively). The critical independent variable was *Studied height-height*, which was coded as 0 for participants the height-voice condition and 1 for those in the height-height condition. Test trials were also coded for *height-voice-conformity* and *height-height-conformity*, with a trial coded as +1 if, on that particular trial for that particular participant, the positive test item conformed to the pattern, and coded as −1 if the negative test item did. The variable $V_1 = V_2$ was used to test whether preference for height-harmonic items was due only to those with identical vowels (e.g., /kidi/). It was +1 if the positive test item had two identical vowels, −1 if the negative test item did, and 0 if both or neither did.[5] The interaction *Studied height-height*$\times$ $V_1 = V_2$ was included to test whether increased preference for height-harmonic items in the *Studied height-height* condition was limited to those with identical vowels.

Two "nuisance variables" were included to model out other sources of variability that had been randomized rather than counterbalanced. Participants in two-alternative forced-choice experiments can be strongly biased towards one response (Yeshurun et al., 2008), so a variable *1st in pair* was included, which was +1 or −1 depending on whether the first 2AFC alternative was the positive or the negative item. Another nuisance variable, $C_1V_1 = C_2V_2$, was used to model bias relating to test items in which the same syllable appeared twice

[4] A very clear and useful exposition of mixed-effects logistic regression can be found in Jaeger (2008).

[5] In this experiment it was never the case that both test items had identical vowels. However, the situation could and did arise in later experiments where height agreement was not a factor.

Table 3: Proportion correct responses in Experiment 1, height-voice vs. height-height. Each cell represents 96 responses (8 from each of 12 participants). The symbols "+" and "−" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying			Familiarized pattern	
+	−	height-voice	height-height	$V_1 = V_2$	height-voice	height-height
tigi	gitu	+	+	+	0.45	0.68
tigu	gitu	+	+	neither	0.60	0.71
tiga	giti	+	–	–	0.60	–
tiga	gitu	+	–	neither	0.66	–
giti	tiga	–	+	+	–	0.61
gitu	tiga	–	+	neither	–	0.68
				MEAN	0.58	0.67

(e.g., [kiki]), which questionnaire responses suggested were highly salient. It was coded as +1 if only the positive item was reduplicated, −1 if only the negative item, and 0 otherwise.[6] Since reduplicated stimuli only occurred in the height-height condition, the interaction *Studied height-height*× $C_1V_1 = C_2V_2$ was also included.

These predictors formed the fixed-effects portion of the model. Because the stimuli were randomly selected and varied from participant to participant, a random intercept was included for each participant. The fixed-effects parameters of the fitted model are shown in Table 4.

Performance in the height-voice group did not differ significantly from chance. The height-height group did significantly better, more than doubling the odds of choosing the pattern-conforming test item ($e^{0.70926} = 2.032$) relative to the height-voice group. Participants showed no significant pre-existing preference for or against height-voice- or height-height-conforming items, indicating that

[6] The procedure described here was followed for all experiments. In Experiment 1, only positive items were ever reduplicated, so $C_1V_1 = C_2V_2$ was never −1. In other experiments, e.g., Experiment 2, all three levels of $C_1V_1 = C_2V_2$ were used.

Table 4: Summary of the fixed effects in the mixed logit model for Experiment 1, height-height vs. height-voice (N = 768, log-likelihood = –483.5)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.22478	0.15394	0.1442
Studied height-height	0.70926	0.26590	0.0076
$V_1 = V_2$	0.09326	0.22928	0.6842
height-height-conforming	–0.10060	0.14732	0.4947
height-voice-conforming	0.09491	0.11140	0.3943
$C_1V_1 = C_2V_2$	–1.19224	0.36010	< 0.001
1st in pair	0.24443	0.07790	0.0017
Studied height-height × $V_1 = V_2$	0.07950	0.33927	0.8147
Studied height-height × $C_1V_1 = C_2V_2$	–0.04948	0.48788	0.9192

the height-height group's preference for height-height-conforming test items was acquired in the experiment. There was no significant preference for or against items with repeated vowels among either the height-voice or the height-height familiarization group; i.e., the superior performance of the height-height group was not restricted to stimuli with identical vowels. There was a very strong aversion to stimuli with a repeated syllable, and a small but highly significant tendency to choose the first test item. Neither $V_1 = V_2$ nor $C_1V_1 = C_2V_2$ interacted significantly with *height-height-conforming.*

These results replicate the main findings of Moreton (2008, Exp. 1). We next discuss the extent to which they can be interpreted as indicative of an inductive bias privileging intra- over interdimensional dependencies.

No alternative explanation for the *Studied height-height* effect is available in terms of biphone or triphone statistics. Every pattern-conforming $V_1C_2V_2$ triphone occurred equally often in the familiarization phase, and every possible $V_1C_2V_2$ triphone occurred equally often in the test. If participants had done the task by simply memorizing the triphones, they would have performed equally well in either familiarization group. If they had used biphones instead, they would have performed better in the height-voice group (in which pattern conformity depends entirely on the V_1C_2 diphone) than in the height-height group (in which all diphones were heard equally often).

Another alternative explanation for the superior salience of the height-height

Table 5: Confusion matrix for consonants.

	Response							
Stimulus	t	d	k	g	Other C	Cluster	No data	Total
C_1 Position								
t	1695	25	20	1	8	0	93	1842
d	63	1471	0	5	6	1	87	1633
k	23	4	1466	89	9	4	92	1687
g	0	25	101	1393	1	7	105	1632
C_2 Position								
t	1647	11	7	1	4	1	28	1699
d	137	1518	2	4	17	0	34	1712
k	6	4	1614	18	1	4	25	1672
g	3	8	124	1529	5	9	33	1711

pattern over the height-voice pattern is that the stimuli were such that height was perceived more accurately than voicing. To test this, 18 participants were sampled from 5 of the experiments reported in this paper, and from another experiment using the same stimuli, paradigm, equipment, and participant pool (not reported here). For each sampled participant, the voice responses from the first and last blocks of the familiarization phase were mixed in random order with those of the other 17 sampled participants from the same experiment. The resulting recording was transcribed by the author or another native American English speaker, in ignorance of the original stimulus which the participant was supposed to be repeating. The transcriptions were then aligned with the corresponding stimuli to derive confusion matrices for the segments in each of the four positions (Tables 5 and 6).[7]

There are two ways in which a difference in featural perceptibility of the stimuli could explain the outcome of the experiment. One is that misperception of the familiarization stimuli might create more perceived nonconformities for

[7] This procedure compares the original stimulus with an experimenter's transcription of the participant's rendition of it. Misperceptions or mispronunciations by the participant may therefore be compounded by misperceptions by the transcriber, thus distorting the picture of the participant's perception. However, the confusions are most likely due to the participant, since the transcriber used a wave editor (Boersma and Weenink, 2010) and could listen to each utterance repeatedly. Misperception by the transcriber is unlikely to directly cancel out misperception by the participant, and so can only lead to an overestimate of the participant's rate of misperception.

Table 6: Confusion matrix for vowels.

	Response								
Stimulus	i/ɪ	u/ʊ	æ	ɔ/a	ɛ	ʌ/ə	Other	No data	Total
V_1 Position									
i	1676	7	0	0	0	1	6	29	1716
u	6	1654	0	0	0	3	7	42	1712
æ	0	3	1100	352	88	4	52	35	1659
ɔ	0	0	47	1497	0	2	103	39	1707
V_2 Position									
i	1635	6	0	0	0	0	0	9	1650
u	9	1561	0	0	0	2	6	15	1593
æ	0	2	898	509	211	140	62	26	1848
ɔ	3	2	44	1439	30	138	27	20	1703

the height-voice group, preventing the pattern from being learned in the first place. A nonconformity would be perceived whenever exactly one of the two critical feature values was perceived as its opposite. The height (high vs. non-high) of V_1 was correctly perceived 99.7% of the time (6433 trials out of 6447, excluding the "Other" and "No data" categories), the height of V_2 98.4% (6620 out of 6729), and the voicing of C_2, 95.5% (6333 out of 6633, excluding the "Other C", "Cluster", and "No data" categories). An height-height-conforming stimulus will thus be misperceived as height-height-nonconforming with probability $0.997(1 - 0.984) + (1 - 0.997)0.984 = 0.019$, whereas an height-voice-conforming stimulus will be misperceived as height-voice-nonconforming with probability $0.997(1 - 0.955) + (1 - 0.997)0.955 = 0.048$, i.e., the height-height group perceived a training corpus which was 98% pattern-conforming, whereas the height-voice group perceived one which was only 95% pattern-conforming. Given the known robustness of lab-learned phonotactics against nonconformities (Chambers et al., 2010), it is unlikely that this small difference in pattern conformity rate was responsible for a twofold difference in odds of a pattern-conforming response.

A second possibility is that misperception of the test items might remove differences in pattern conformity, and do so more in the height-voice than the height-height group. For this to happen, exactly one of the two test items in a pair has to be misperceived in a conformity-reversing way (the chance that it will happen to both items is negligible). That probability is $2 \cdot (1 - 0.19) \cdot 0.19 = 0.037$ for height-height-conformity, and $2 \cdot (1 - 0.048) \cdot 0.048 = 0.091$ for height-voice-conformity; i.e., an extra 5.4% of test trials are ruined by misperception in the height-voice condition. This again is not enough to explain the near-doubling of the odds of a pattern-conforming response in the height-height condition relative to the height-voice condition.

3. Experiment 2

The height-height and height-voice patterns of Experiment 1 differ in another way as well: Both of the critical segments of the height-voice pattern are buried in the middle of the stimulus, whereas the height-height pattern involves a word-final segment. Segments at the beginning and end of a word are known to be especially salient (Endress and Mehler, 2010). Perhaps the height-voice pattern is harder to acquire simply because the critical segments are inconspicuous.

To test whether the salience of patterns involving word edges is enough to explain the higher level of performance in the height-height condition of Experiment 1, Experiment 2 compared the height-voice pattern of Experiment 1 with a "voice ... height" pattern linking word-initial C_1 with word-final V_2. If the results of Experiment 1 are due to edge salience, then a similar result should be found in Experiemnt 2: The non-adjacent dependency between two edge segments should elicit better performance than the adjacent dependency between two medial segments.

3.1. Method

The same set of 256 $C_1V_1C_2V_2$ stimuli was used. As in Experiment 1, a stimulus was defined as height-voice-conforming when V_1 was high if and only if C_2 was voiced, and as voice ... height-conforming when V_2 was high if and only if C_1 was voiced. For each participant, a familiarization set of 32 pattern-conforming stimuli was randomly selected, subject to the constraint that each of the 8 pattern-conforming critical diphones (V_1C_2 or $C_1 \ldots V_2$, depending on familiarization group) had to occur equally often. For the test set, 64 novel items were randomly chosen and arranged into pairs subject to the condition that half of the pairs matched a stimulus that conformed to both patterns with one that conformed to neither, while the other half matched a stimulus that conformed to only one pattern with one that conformed only to the other. Twelve test sets were generated, each heard by one participant in each familiarization group. In other respects, the procedure was like that of Experiment 1.

Twenty-four volunteers, recruited as for Experiment 1, participated in this experiment. The height-voice group consisted of 7 female and 5 male participants, with a mean age of 22.3 years (s.d. 3.7 years). The voice ... height group consisted of 8 female and 4 male participants, with a mean age of 25.6 years (s.d. 9.8 years). Participant gender had no main effect nor interaction with any of the other variables in Experiment 1. Since women outnumber men two to one in the population from which participants were recruited, no effort was made to equalize the numbers of male and female participants in this or later experiments.

3.2. Results

The raw response proportions are shown in Table 7. The results were analyzed in a mixed-effects logistic-regression model, as in Experiment 1. The fixed-effects portion of the model is shown in Table 8.

Participants in the height-voice group chose the pattern-conforming test item significantly more often than chance. Performance in the voice ... height group was numerically smaller, but the difference was only marginally significant There were no significant effects of pre-existing preference for height-voice- or voice ... height-conformity. Participants again showed a significant tendency to reject items with a repeated syllable, and to prefer the first of the two response options. (Unlike in Experiment 1, there was no need to test for an interaction between *Studied voice ... height* and $C_1V_1 = C_2V_2$, since stimuli with repeated syllables were equally likely in either familiarization condition.)

These results provide no support for the edge-salience interpretation of Ex-

Table 7: Proportion correct responses in Experiment 2, voice . . . height vs. height-voice. Each cell represents 192 responses (16 from each of 12 participants). The symbols "+" and "–" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying		Familiarized pattern	
+	–	height-voice	voice . . . height	height-voice	voice . . . height
tiga	gita	+	+	0.61	0.53
tigu	gitu	+	–	0.55	–
gitu	tigu	–	+	–	0.54
			MEAN	0.58	0.54

Table 8: Summary of the fixed effects in the mixed logit model for Experiment 2, voice . . . height vs. height-voice (N = 768, log-likelihood = –512.1)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.38154	0.15060	0.0113
Studied voice . . . height	–0.35707	0.21522	0.0971
voice . . . height-conforming	0.11399	0.10555	0.2802
height-voice-conforming	–0.03287	0.10432	0.7527
$C_1V_1 = C_2V_2$	–0.85643	0.24558	<0.001
1st in pair	0.25492	0.07444	<0.001

periment 1. The predicted advantage for the voice ... height pattern was not found (if anything, the trend was in the opposite direction).

4. Experiment 3

If the height-height advantage in Experiment 1 was due to generally greater salience of intradimensional dependencies, then the dimensions themselves should not matter. To test this hypothesis, Experiment 1 was repeated with backness substituting for height. A stimulus was *backness-backness-conforming* when V_1 and V_2 were both back (/u ɔ/) or both front (/i æ/). It was *backness-voice-conforming* when V_1 was back if and only if C_2 was voiced.

4.1. Method

Twenty-four volunteers participated in this experiment. The backness-voice group consisted of 11 female and 1 male participants, with a mean age of 32.2 years (s.d. 15.1 years). The backness-backness group consisted of 10 female and 2 male participants, with a mean age of 30.3 years (s.d. 15.9 years). In all other respects, the experiment was identical to Experiment 1.

4.2. Results

Table 9 shows the proportion of pattern-conforming responses in each condition. The results were analyzed statistically in the same way as those of Experiment 1. The fixed-effects part of the model is shown in Table 10.

The results are very similar to those of Experiment 1. Performance in the backness-voice condition was not significantly different from chance, whereas familiarization on the backness-backness pattern significantly increased the odds of a pattern-conforming response. No pre-existing preference for either pattern was found, nor was there a preference for items with repeated vowels in either familiarization condition. The effects of a repeated vowel or syllable did not differ between the two familiarization conditions. Experiment 3 thus replicates the finding of Experiment 1 that an intradimensional biconditional is learned better than an interdimensional one.

Table 9: Proportion correct responses in Experiment 3, backness-voice vs. backness-backness. Each cell represents 96 responses (8 from each of 12 participants). The symbols "+" and "–" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying			Familiarized pattern	
+	–	backness-voice	backness-backness	$V_1 = V_2$	backness-voice	backness-backness
taga	tikæ	+	+	+	0.36	0.55
tagu	tikæ	+	+	neither	0.48	0.60
tagæ	gækæ	+	–	–	0.55	–
tagæ	gikæ	+	–	neither	0.49	–
gækæ	tagæ	–	+	+	–	0.66
gikæ	tagæ	–	+	neither	–	0.64
				MEAN	0.47	0.61

Table 10: Summary of the fixed effects in the mixed logit model for Experiment 3, backness-backness vs. backness-voice (N = 768, log-likelihood = –502)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.02454	0.15021	0.8702
Studied backness-backness	0.53578	0.25574	0.0362
$V_1 = V_2$	–0.04589	0.22495	0.8384
backness-backness-conforming	–0.00916	0.14518	0.9497
backness-voice-conforming	–0.14114	0.10712	0.1876
$C_1V_1 = C_2V_2$	–1.36553	0.37973	<0.001
1st in pair	0.16233	0.07556	0.0316
Studied backness-backness×$V_1 = V_2$	0.20502	0.32631	0.5298
Studied backness-backness×$C_1V_1 = C_2V_2$	0.22509	0.50404	0.6552

This experiment also reverses Experiment 1's confound between pattern-learning and featural perceptibility. Table 6 shows that backness was correctly perceived in V_1 on 93.3% of trials (6018 of 6447, excluding the "Other" and "No data" categories), whereas in V_2 it was correctly perceived on just 87.4% (5884 of 6729). Participants in the backness-backness condition perceived nonconforming stimuli on about $0.933 \cdot (1 - 0.874) + (1 - 0.933) \cdot 0.874) = 17.6\%$ of familiarization trials, whereas those in the backness-voice condition did so on only $0.933 \cdot (1 - 0.955) + (1 - 0.933) \cdot (0.955) = 10.6\%$. Thus, in Experiment 3 the better-learned pattern is the less-perceptible one, whereas in Experiment 1, the reverse is true. This rules out featural perceptibility as an alternative explanation for the results of Experiment 1.

5. Experiment 4

Experiments 1 and 3 found a learning advantage for intradimensional dependencies over interdimensional ones. However, the intradimensional dependencies related two vowels, whereas the interdimensional dependencies related a vowel and a consonant. Experiments with non-phonological stimuli, and even with non-human subjects, have repeatedly shown that perceptual similarity facilitates association between the elements of a compound stimulus (Köhler, 1941; Prentice and Asch, 1958; Arnold and Bower, 1972; Rescorla and Gillan, 1980; Rescorla, 1986; Creel et al., 2004; Rescorla, 2008). To test whether this effect is sufficient to explain the results of Experiments 1 and 3, Experiment 4 compared two interdimensional patterns, one relating vowels and the other relating a vowel and a consonant. This experiment is similar to Experiment 1, except that the height-height dependency is replaced by a height-backness dependency: In pattern-conforming stimuli, V_1 is high if and only if V_2 is back.

5.1. Method

Twenty-six participants were recruited for this experiment. Results from two participants were excluded from analysis (one due to conspicuous difficulty hearing the stimuli; the other due to experimenter error). Among the remaining

Table 11: Proportion correct responses in Experiment 4, height-backness vs. height-voice. Each cell represents 192 responses (16 from each of 12 participants). The symbols "+" and "−" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying		Familiarized pattern	
+	−	height-voice	height-backness	height-voice	height-backness
kati	kitæ	+	+	0.60	0.61
katæ	gita	+	−	0.53	−
gita	katæ	−	+	−	0.62
			MEAN	0.57	0.61

participants, the height-voice group consisted of 11 female and 1 male volunteer with a mean age of 22.7 years (s.d. 6.2 years), while the height-backness group consisted of 9 female and 3 male volunteers with a mean age of 23.2 years (s.d. 5.3 years). This experiment did not otherwise differ from Experiment 1.

5.2. Results

Raw response probabilities are shown in Table 11, and the fixed-effects part of the statistical model in Table 12. As in the previous experiments, participants strongly dispreferred stimuli with a repeated syllable, and tended to choose the first of the two responses on each trial. The intercept narrowly missed significance at the usual 5% level (est. = 0.30757, s.e. = 0.15744, z=1.954, p = 0.0508), indicating that performance in the height-voice group was above chance with marginal confidence. No pre-existing preference was found for height-voice- or height-backness-conforming response options. These results are similar in terms of both magnitude and precision to those found in Experiment 1, However, this time performance in the height-backness condition was no better than that in the height-voice control group.

If the positive results of Experiments 1 and 3 had been due to a Gestalt principle which grouped the vowels together perceptually, this experiment ought to have gotten a positive result as well. The failure to find one tells against the

Table 12: Summary of the fixed effects in the mixed logit model for Experiment 4, height-backness vs. height-voice (N = 768, log-likelihood = –492)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.30757	0.15744	0.0508
Studied height-backness	0.02335	0.22323	0.9170
height-backness-conforming	0.16638	0.10652	0.1183
height-voice-conforming	-0.05326	0.10859	0.6238
$V_1 = V_2$	-0.35293	0.13855	0.0109
$C_1V_1 = C_2V_2$	-0.77386	0.24931	0.0019
1st in pair	0.38055	0.07774	< 0.001

Gestalt explanation. This result cannot be attributed to a general difficulty in learning dependencies involving height, or dependencies involving backness, since Experiments 1 and 3 showed that height-height and backness-backness dependencies were learned better than controls. Rather, the within-tier repetition of a single feature seems to enjoy a special advantage over other within-tier dependencies.

6. Experiments 5 and 6

The within-tier conditions in Experiments 1, 3, and 4 were restricted to dependencies between the two vowels. However, several recent experiments have uncovered evidence that biconditionals relating vowels may be easier to learn than analogous dependencies between consonants (Toro et al., 2008,?; Pons and Toro, 2010; Nevins, 2010), raising the possibility that the intradimensional advantage observed in Experiments 1–4 might be peculiar to the vocalic tier, or peculiarly strong there. Some contrary evidence was found by Wilson (2003), who observed an intradimensional advantage in consonant-consonant patterns, but that experiment had no vowel-vowel analogue for comparison.

Consonantal analogues were constructed for the vowel-dependency experiments. Experiment 5 was like Experiment 1, except that agreement between the height of the two vowels was replaced with agreement between the voicing of the two consonants. Experiment 6 was like Experiment 4, except that the height-backness dependency between the vowels was replaced with a place-voice dependency between the consonants (C_1 was coronal if and only if C_2 was voiced).

6.1. Method

Twenty-six participants were recruited for Experiment 5. Data from two participants was excluded because they specified languages other than English as their first language on the post-experiment questionnaire. The final height-voice group consisted of 9 women and 3 men with a mean age of 24.3 years (s.d. 8.4 years), while the voice-voice group consisted of 7 women and 5 men with a mean age of 24.3 years (s.d. 2.7 years). Twenty-four volunteers participated in Experiment 6. Both groups consisted of 8 women and 4 men. The mean age in the place-voice group was 21.1 years (s.d. 6.0 years); that in the height-voice group was 19.6 years (s.d. 1.6 years). The experimental procedure was identical to that of Experiment 1.

6.2. Results

The raw response proportions for Experiment 5 (the analogue of Experiment 1) are shown in Table 13, and the logistic-regression results in Table 14. Aside from the nuisance variables $C_1V_1 = C_2V_2$ and *1st in pair*, none of the fitted coefficients was significantly different from zero. Participants in the height-voice condition performed marginally above chance, while those in the voice-voice condition performed marginally better than those in the height-voice condition. There were no significant pre-existing preferences or dispreferences for stimuli conforming to either pattern, or for stimuli in which the same consonant occurred twice. The effects of a repeated consonant or syllable did not differ significantly between the voice-voice and height-voice conditions (though the magnitudes of the two interaction terms were numerically quite large).

The results and analysis of Experiment 6 are shown in Tables 15 and 16. The outcome here was similar to that of Experiment 4. Participants in the height-voice condition performed significantly above chance, and performance in the place-voicecondition was not significantly different from that in the height-voice

Table 13: Proportion correct responses in Experiment 5, height-voice vs. voice-voice. Each cell represents 96 responses (8 from each of 12 participants). The symbols "+" and "–" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying			Familiarized pattern	
+	–	height-voice	voice-voice	$C_1 = C_2$	height-voice	voice-voice
didu	gutu	+	+	+	0.45	0.51
gidu	gutu	+	+	neither	0.55	0.68
kidu	dadu	+	–	–	0.67	–
kidu	gadu	+	–	neither	0.61	–
dadu	kidu	–	+	+	–	0.55
gadu	kidu	–	+	neither	–	0.65
				MEAN	0.57	0.60

Table 14: Summary of the fixed effects in the mixed logit model for Experiment 5, voice-voice vs. height-voice (N = 768, log-likelihood = -497.9)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.33151	0.17188	0.0538
Studied voice-voice	0.49761	0.28561	0.0815
voice-voice-conforming	–0.13096	0.14863	0.3783
height-voice-conforming	0.00419	0.10713	0.9688
$C_1 = C_2$	–0.08042	0.22526	0.7211
$C_1V_1 = C_2V_2$	–1.26143	0.42226	0.0028
1st in pair	0.28681	0.07597	0.0002
Studied voice-voice × $C_1 = C_2$	–0.34694	0.32116	0.2800
Studied voice-voice × $C_1V_1 = C_2V_2$	0.67560	0.55351	0.2223

Table 15: Proportion correct responses in Experiment 6, place-voice vs. height-voice. Each cell represents 192 responses (16 from each of 12 participants). The symbols "+" and "−" refer to the positive and negative (pattern-conforming and pattern-nonconforming) response options on each forced-choice trial. Some stimulus pairs appear twice because they are coded differently in the two familiarization conditions.

Stimulus pair (e.g.)		Stimulus satisfying		Familiarized pattern	
+	−	height-voice	place-voice	height-voice	place-voice
tigu	tiku	+	+	0.51	0.53
tæku	tægu	+	−	0.60	−
tægu	tæku	−	+	−	0.56
			MEAN	0.55	0.55

Table 16: Summary of the fixed effects in the mixed logit model for Experiment 6, place-voice vs. height-voice (N = 768, log-likelihood = -502.6)

Coefficient	Estimate	SE	$Pr(>\lvert z\rvert)$
Intercept	0.30508	0.15350	0.0469
Studied place-voice	0.08897	0.21521	0.6793
$C_1 = C_2$	-0.22197	0.13460	0.0991
place-voice-conforming	-0.19646	0.10644	0.0649
height-voice-conforming	-0.04524	0.10580	0.6690
$C_1V_1 = C_2V_2$	-1.31380	0.29461	< 0.001
1st in pair	0.26189	0.07541	0.0005

condition. There were no significant pre-existing preferences for height-voice- or place-voice-conforming stimuli.

To compare these results with those from the vowel experiments, results of the four corresponding experiments were analyzed together (Experiment 3 was omitted because it had no consonantal mate). The model included the contrast terms shown in Table 17, fully crossed with a new variable, *Final Triphone*, which had the value +1 for Experiments 1–4 and -1 for Experiments 5–6. The *Final Triphone* variable thus distinguished the "vowel experiments", in which familiarization presented all legal final triphones, from the "consonant experiments", in which it presented all legal initial ones. The model again included the nuisance variables $C_1V_1 = C_2V_2$ and *1st in pair*, which had had

Table 17: Orthogonal contrasts for the combined analysis of Experiments 1, 4, 5, and 6.

		Contrast		
Experiment		*One Tier*	*One Feature*	*Final Triphone*
Vowel experiments				
1	height-height	1	1	1
	height-voice	-1	0	1
4	height-backness	1	-1	1
	height-voice	-1	0	1
Consonant experiments				
5	voice-voice	1	1	-1
	height-voice	-1	0	-1
6	place-voice	1	-1	-1
	height-voice	-1	0	-1

Table 18: Logistic-regression analysis of Experiments 1, 4, 5, and 6, parametrized by the orthogonal contrasts shown in Table 17 ($N = 3072$, log-likelihood = –1993).)

Coefficient	Estimate	SE	$Pr(>\|z\|)$
Intercept	0.41217	0.04035	<0.001
One Tier	0.12467	0.04006	0.0019
One Feature	0.19351	0.05734	<0.001
Final Triphone	0.08286	0.03988	0.0377
One Tier × *Final Triphone*	0.08369	0.03992	0.0360
One Feature × *Final Triphone*	0.02001	0.05690	0.7251
$C_1V_1 = C_2V_2$	-1.19109	0.11737	< 0.001
1st in pair	0.28577	0.03801	< 0.001

highly significant effects in all previous analyses, and excluded the nuisance variables and interactions which had not had significant effects in the individual-experiment analyses. This analysis gains a bit of statistical power, relative to the individual analyses, by collapsing together the height-voice conditions within each pair of experiments. The results are shown in Table 18.

When the four experiments are taken together, the highly significant effect of *One Feature*, and the lack of a significant interaction between *One Feature* and *Final Triphone*, confirms that the intradimensional dependencies (height-height and voice-voice) were learned better than the corresponding interdimensional ones (height-backness and place-voice).[8]

7. General discussion

The results of these experiments indicate that a phonotactic pattern based on agreement between two instances of the same phonological feature is easier to learn than one based on correlation between instances of two different features. This is true even though the segments participating in the inter-dimensional dependency were adjacent, while the ones in the intra-dimensional dependency were not. Further experiments led us to reject alternative explanations based on sensitivity to repetitions of the exact same segment, to consonant-consonant or vowel-vowel dependencies in general, and to salient word-edge segments. Phonotactic learning thus exhibits the same intra-dimensional advantage that has been found in non-linguistic concept learning and in natural-language typology.[9]

7.1. Agreement and other intra-dimensional dependencies

Six different intra-dimensional dependencies can be defined using two binary features, as shown in Table 19. The experiments in this paper addressed only one of them, namely identity between two feature instances ($[\alpha F_1] \ldots [\alpha F_2]$, the first of the biconditionals in Table 19). Those are artificial analogues of the very common natural-language featural agreement (harmony) patterns. If the $+$ and $-$ feature values are interpreted as Boolean truth values, then these patterns express the if-and-only-if (IFF) relation.[10]

[8] The results do not change materially if the omitted nuisance variables (*height-height-conforming*, $V_1 = V_2$, etc.) are included. The effect of *One Feature* becomes slightly larger and more significant.

[9] These results are consistent with the previous findings reviewed in §1.2, except for Kuo (2009)'s Experiments 1 and 2, which found no significant difference between a place-place pattern and an aspiration-place pattern. There was some indication in Kuo's data of a slight advantage for the place-place pattern; in particular, when compared to a third, much more difficult place-aspiration-place pattern (Experiment 3), participant responses in the early blocks of the place-place pattern were significantly more pattern-conforming, whereas those of the aspiration-place pattern were not. Effect strength may also have been attenuated by a distractor task which separated familiarization from testing. However, the inter- vs. intra-dimensional effect, if present at all, remains surprisingly weak compared to other effects found in the same study.

[10] Since the two features are both instances of the same feature, it does not matter whether we count $+$ as TRUE and $-$ as FALSE, or vice versa; an identity pattern always translates

Table 19: The possible dependencies defined over two binary features. Each dependency is defined by marking pattern-conforming feature combinations with ●, and non-conforming ones with ○. The letters "A" and "D" indicate the dependencies that would be assimilations or dissimilations if F_1 and F_2 are instances of the same feature (i.e., in the intra-dimensional case).

	F_2											
F_1	+	−	+	−	+	−	+	−	+	−	+	−
+	●	○	●	●	○	●	●	●	●	○	○	●
−	●	●	○	●	●	●	●	○	○	●	●	○
	$+\rightarrow+$ A		$-\rightarrow-$ A		$+\rightarrow-$ D		$-\rightarrow+$ D		$+\leftrightarrow+$ A		$+\leftrightarrow-$ D	
	Conditionals								Biconditionals			

The experiments in this paper did not address the complementary anti-identity patterns ($[\alpha F_1] \ldots [-\alpha F_2]$), which are logically interpretable as exclusive-or (XOR). This is a significant omission, since dissimilation patterns in general seem to be much rarer in natural language than assimilation both synchronically (Bye, 2011) and diachronically (Campbell, 2004, 30), which might be a sign that inductive bias favors agreement in particular rather than intra-dimensional patterns in general. Other artificial-phonology studies, however, have tested for learnability differences between matched harmony and anti-harmony patterns, and have not found any (Pycha et al. 2003; Wilson 2003; Koo and Cole 2006; see also Kuo 2009, Experiment 1, and Skoruppa and Peperkamp 2011). This suggests that if there is a phonological harmony bias, it is a small relative to the general advantage for intra-dimensional dependencies. I do not know of any analogous non-linguistic experiments that directly compare IFF with XOR intra-dimensionally. A large IFF advantage in non-linguistic learning would be evidence for domain-specific processes in phonological learning.

The present experiments also ignored the intra-dimensional conditionals

as IFF and an anti-identity pattern as XOR. For dependencies defined over instances of two different features, it does matter which value of the feature is chosen as TRUE. For example, "voiced if and only if high" and "voiceless unless (i.e., XOR) high" describe the same pattern but assign different truth values to the voicing feature (voiced = TRUE in the former, voiceless = TRUE in the latter). The IFF/XOR distinction has no meaning for inter-dimensional biconditionals unless particular feature values can be recognized as primary on some principled basis such as markedness, frequency, or perceptual salience.

shown in Table 19, which are common in natural-language assimilation and dissimilation patterns. For example, “Lyman’s Law” in Japanese bans the occurrence of two voiced obstruents in a stem, but allows two voiceless obstruents or one voiced and one voiceless obstruent, in either order (Itô and Mester, 1995). Lyman’s Law can be formulated as a conditional (“If the first of two obstruents is [+voiced], the second must be [−voiced]”), or equivalently as a disjunction (“Either the first or the second of two obstruents must be [−voiced]”). Since an intra-dimensional conditional includes the biconditional as a subcase, any apparent intra-dimensional advantage among conditionals would need to be examined carefully to check whether the entire effect was due to the biconditional stimuli alone.

7.2. Verbal complexity

Previous accounts of the intra-dimensional advantage in general concept learning have attributed it to the complexity of intra-dimensional patterns when stated verbally. In the studies of Ciborowski and Cole (1973), the authors told participants beforehand that their goal would be “to find out the rule”, and asked them immediately after each problem to explain on what basis they had sorted the cards. The stimulus properties made the rules easy to verbalize. The intra-dimensional target rules could be verbally abbreviated in ways that the inter-dimensional ones could not; e.g., “red and red” could be expressed as “two red” or “both red”, whereas there was no shorter form for “red and triangle”. Learning was significantly better for intra-dimensional problems when the short form was reported. An intra-dimensional advantage associated with short-form rule reports was also found with boys and young men of all educational levels in rural Liberia as long as previously-classified stimulus cards were left face-up. When this memory aid was not available, participants were much less likely to verbalize any rule, and the intra-dimensional advantage disappeared. This interpretation is consistent with the negative results of Shepard et al. (1961), whose Experiments II and III found no difference in verbalization complexity between otherwise comparable inter-and intra-dimensional relations, and, concomitantly, no difference in classification performance. [11]

However, phonological features are very difficult for naïve participants to verbalize. In the present study, responses on the post-experiment questionnaire almost never correctly distinguished pattern-conforming stimuli from non-conforming ones. Most participants either disclaimed knowledge of a pattern, or wrote a statement which was equally valid for both positive and negative stimuli (e.g., "They were all two syllables and all began with hard consonants such as k, t, g, d, etc."). A handful did describe the pattern more or less correctly, but their data was excluded from the analysis and replaced. Learning in these experiments therefore appears to have been wholly implicit, and participants would not have benefited from the availability of verbal predicates such as "both" or "same" in intra-dimensional conditions. Even in non-linguistic learning, verbalization is not a necessary precondition for the anomalous ease of the intra-dimensional equality relation, compared to conjunction or disjunction (Hunt and Hovland, 1960).

The shared intra-dimensional advantage is therefore not due to a domain-general learning mechanism that is sensitive to verbal complexity. On the other hand, non-verbal models of domain-general learning make no distinction between intra- and inter-dimensional dependencies (Gluck and Bower, 1988; Anderson, 1991; Kruschke, 1992; Nosofsky et al., 1994; Love et al., 2004; Feldman, 2006), and so leave both the phonological and the non-phonological intra-dimensional superiority effects unexplained.

7.3. Non-verbal complexity

An alternative account is that the complexity hypothesis is essentially correct, but applies to an implicit, non-verbal level of representation — that the "vocabulary" of the learner's hypothesis space contains special predicates that facilitate detection of featural agreement by, in effect, giving the learner addi-

[11] It is also consistent with the observation that the experiments of Laughlin and Jordan (1967) and Laughlin (1968), which found an intra-dimensional biconditional to be anomalously easy compared to conjunction and disjunction, required participants to verbalize a hypothesis on every trial.

tional means to detect it. This proposal is introduced, and discussed in more detail, in Pater and Moreton (unpublished); what follows is a précis.

Consider a stimulus space defined by three binary features, F_1, G, and F_2, corresponding to, e.g., the height of V_1, the voicing of C_2, and the height of V_2. Suppose a learner is equipped with the following predicates:

$$\begin{array}{cc}
(F_1, F_2) = (+,+) & (F_1, G) = (+,+) \\
(F_1, F_2) = (+,-) & (F_1, G) = (+,-) \\
(F_1, F_2) = (-,+) & (F_1, G) = (-,+) \\
(F_1, F_2) = (-,-) & (F_1, G) = (-,-) \\
(F_1, F_2) = (\alpha,\alpha) &
\end{array}$$

The learner has only one way to represent a positive correlation between F_1 and G, viz., $(F_1, G) = (+,+) \vee (F_1, G) = (-,-)$. Each half of the disjunction must be learned piecemeal; the learner has no way of using evidence about $(+,+)$ to make inferences about $(-,-)$. A positive correlation between F_1 and F_2, however, can be learned in two ways, either piecemeal, as the disjunction $(F_1, F_2) = (+,+) \vee (F_1, F_2) = (-,-)$, or wholesale, as the single predicate $(F_1, F_2) = (\alpha,\alpha)$. This can provide a learning advantage for the intradimensional correlation over the interdimensional one, even without a hard-wired bias favoring shorter formulas.

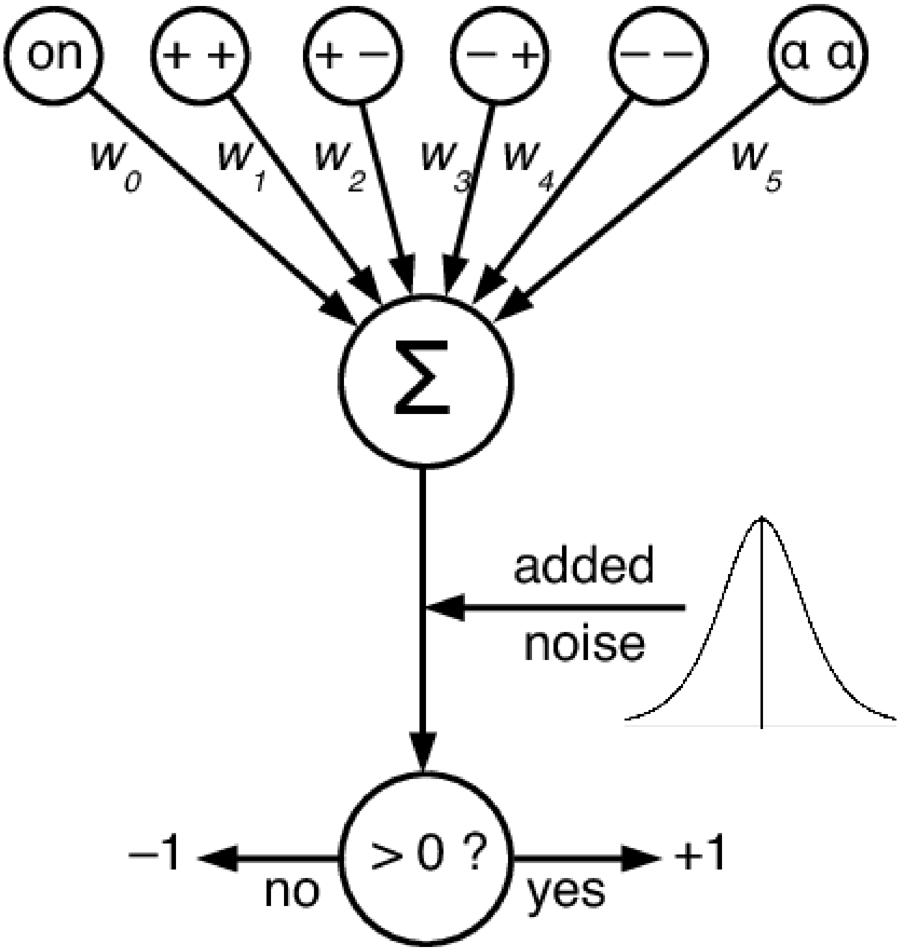
The effect can be demonstrated with the aid of a learning model which has figured prominently in accounts of domain-general learning (Rosenblatt, 1962; Rescorla and Wagner, 1972; Gluck and Bower, 1988; Kruschke, 1992; Love et al., 2004), and which has close analogues in the phonology literature (Boersma, 1997; Boersma and Hayes, 2001; Jäger, 2007; Pater, 2008; Magri, 2008; Boersma and Pater, 2008): the single-layer feed-forward neural net, or perceptron. In the model shown in Figure 1, there are four input units whose receptive fields correspond to the four possible combinations of two binary features; e.g., the $+-$ unit outputs $+1$ when the two stimulus features are $(+,-)$, and -1 otherwise. There is also a bias unit that is always "on", with constant output $+1$, to accommodate base-rate effects. The input units correspond to the descriptive primitives available to the learner. This network's architecture is based on that of the Configural Cue Model of Gluck and Bower (1988), but has been simplified by omitting four single-feature input units, which are irrelevant here.

In the "HV" (height-voice) condition, these are the only input units. In the "HH" (height-height) condition, there is an additional unit which outputs $+1$ for the stimuli $(+,+)$ and $(-,-)$, and -1 for the other two stimuli. Logistic noise with zero mean and unit variance is added to the weighted sum of the input units' activation to simulate human nondeteriminism. (Logistic noise was used in preference to Gaussian because it is mathematically more tractable; see Appendix.) The network's output is $+1$ or -1 depending on whether the weighted sum of the input activations is positive or not (Figure 1).

The weight vector $\mathbf{w}$ was initialized to the zero vector, and the network was trained for 5000 trials using the perceptron learning rule, $\mathbf{w} \leftarrow \mathbf{w} + \eta(t-o)\mathbf{i}$, where t and o were the correct and actual outputs on that trial ($+1$ or -1), $\mathbf{i}$ the vector of input-unit activations (each $+1$ or -1), and $\eta = 0.001$ was the learning rate. All four training stimuli were presented with equal probability.

As Figure 2 shows, learning proved to be faster in the height-height condition than in the height-voice condition (theoretically, exactly twice as fast; see Appendix for an analysis). This happened because the HH condition afforded more numerous and more general cues. In both conditions, the net learned by acquiring positive weights on the input nodes that responded to the stimuli $(+,+)$ and $(-,-)$, and negative weights on those that responded to $(+,-)$ and $(-,+)$. In the HV condition, that meant up-weighting the $++$ and $--$ nodes, and down-weighting the $+-$ and $-+$ nodes. This happened piecemeal; the connection to, e.g., $++$ could only be strengthened on trials where the stimulus was $(+,+)$. In the HH condition, however, there was an additional input unit $\alpha\alpha$ which responded positively to both of the positive stimuli and negatively to both of the negative stimuli. The extra unit not only provided extra activation, it also gained connection strength faster, since it was up-weighted every time either $++$ or $--$ was up-weighted. The net thus used the $\alpha\alpha$ input unit to generalize, learning about the correct classification of $(+,+)$ from experience

Figure 1: Single-layer perceptron model of height-height pattern learning. The height-voice learner lacks the "$\alpha\alpha$" unit.



with $(-,-)$, and vice versa. The symmetry of the network allows it to learn the inverse pattern (disharmony) in the same way and at the same rate, with only the signs of the weights reversed.[12]

This account of the intra-dimensional superiority effect depends crucially on both the existence of predicates of the form "$(F_1, F_2) = (\alpha, \alpha)$", and the absence of those of the form "$(F_1, G) = (\alpha, \alpha)$". It is thus consistent with previous hypotheses that there exist mental symbolic variables (Marcus et al., 1999), that such variables can range over phonological features (Halle, 1962; Bach and Harms, 1972), and that there are specialized mental predicates for within-stimulus identity in general (Endress et al., 2007; Endress and Mehler, 2010) and for within-stimulus featural identity in particular (Goldsmith, 1976; Rose and Walker, 2004). The similar inductive biases in phonological and non-phonological learning emerge, in this proposal, from the application of a domain-general learning algorithm (the perceptron rule) on a domain-specific set of predicates (constraints stated over phonological representations) which are subject to a domain-general restriction (that variables only relate instances of the same feature).[13]

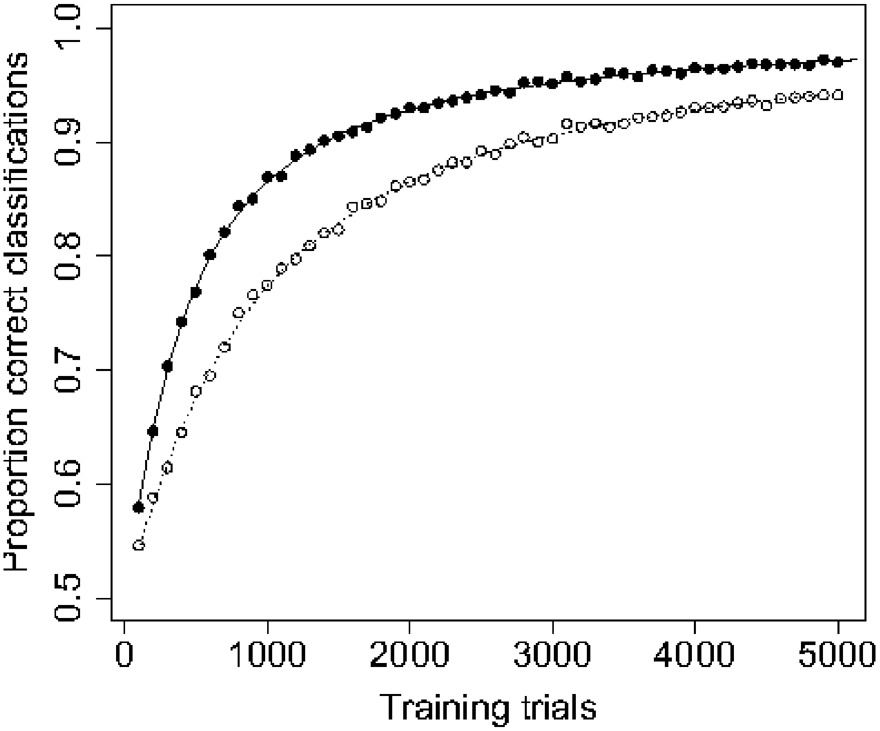
7.4. Summary and conclusions

As usual in lab studies, there is no way to guarantee in advance that the phenomenon we are studying in the lab is the same one that we see in nature,

[12] The well-known inability of perceptrons to learn IFF and XOR relations applies only to those in which each input unit corresponds to a single feature (Minsky and Papert, 1969). The HH/HV network, like the Configural Cue Model it is based on, circumvents this limitation by using input units which correspond to combinations of feature values.

[13] A simple linear theory of phonetic features, essentially that of Chomsky and Halle (1968), has proven adequate for both the empirical and the modelling purposes of the present study. Some phonological theories propose highly structured representations in which features are independent entities which can be associated with serial positions in a one-to-many or many-to-one relationship (Leben, 1973; Goldsmith, 1976; Clements, 1985; McCarthy, 1986, 1988). A major aim of this approach is to distinguish intra-dimensional patterns from inter-dimensional ones and to favor the former by providing grammatical predicates for encoding them. No predicates are specified for encoding inter-dimensional patterns, but since languages have them, the theory must provide some more-general schema to accommodate them. Once this is done, the model is similar to the one sketched in this section: There is a general schema which can represent both kinds of pattern, and a special one for intra-dimensional patterns alone. Learning models incorporating autosegmental representations are a recent development (Hayes et al., 2008; Heinz et al., 2011).

Figure 2: Simulation results. Learning of height-height and height-voice patterns by the perceptron model, showing the model's probability of correct classification as a function of the number of training trials. Dots show empirical average of 100 replications in each condition. Lines show analytic predictions (see Appendix).



so we really have to do with three domains: non-linguistic categories, artificial phonology, and natural-language phonology. The main possibilities are (1) all three kinds of category are learned using the same cognitive mechanism; (2a) natural and artificial phonology are learned using a dedicated mechanism; (2b) only natural-language phonology is learned using a dedicated mechanism. One way to distinguish among these three hypotheses is to look for contingent properties which are shared across domains, or are shared in two domains to the exclusion of the third. (Some other ways are discussed in Moreton and Pater to appear.)

The present experiments show that phonological and non-linguistic learning share a contingent inductive bias, i.e., one which a reasonable learning algorithm could in principle lack (and which existing models of non-linguistic category learning in fact do lack). The intra-dimensional learning advantage is also matched by an asymmetry in the cross-linguistic frequencies of phonotactic patterns. As noted above (p. 3), all three domains likewise share an advantage for patterns in which fewer features are relevant. The hypothesis that best explains this particular set of facts is (1), since either of the other two would make the parallels coincidental.

It might be objected that intra-dimensional and featural-simplicity biases are so obvious and generic as to be inevitable in any learner, and hence that the sharing of them does not tell against Hypotheses (2a) and (2b). The intra-dimensional bias cannot be entirely obvious or inevitable if so many non-linguistic learning models leave it out, and models have also been seriously considered which reverse the featural-simplicity bias by acquiring biconditionals faster than single-feature affirmations (Neisser and Weene, 1962).[14] However, it would be premature to conclude that Hypothesis (1) is correct, since we still have only these few points of comparison across the three domains. The is-

[14] The reason for this reversal, the authors suggested, was that learning success was defined as elimination of all competing hypotheses (rather than as reaching a particular performance criterion), and because the affirmation was in a more-densely-populated region of their hypothesis space than the biconditional and hence had more competitors (Neisser and Weene, 1962, 644). The authors note that the model is unrealistic as an account of human behavior.

sue can only be settled by systematic comparison of isomorphic patterns across domains: inductive biases in linguistic and non-linguistic learning, typological frequencies in natural-language phonology, and the phonetic interactions out of which phonological patterns develop.

The present results also support the specific hypothesis that the high frequency of featural-agreement patterns in natural language is due at least in part to an inductive bias that makes them especially easy to learn or innovate. To reject that hypothesis would require us to believe that *only* natural-language phonological learning lacks an intra-dimensional advantage, and that agreement patterns are frequent for some other reason. The most plausible "other reason" is that intra-dimensional dependencies are based on stronger phonetic interactions than inter-dimensional ones (see above, p. 5); however, this phonetic claim has not been supported when the relevant comparisons have been made (Moreton 2008, 2010; see Kapatsinski 2011 for a possible alternative). Consequently, the most plausible interpretation of the results — that inductive bias contributes to the high frequency of agreement patterns — supports the hypothesis that natural languages are adapted to, and informative about, the inductive biases of the learner, whether domain-specific or domain-general (e.g. Chomsky and Halle, 1968; Prince and Smolensky, 1993; Archangeli and Pulleyblank, 1994; Saffran, 2002; Newport and Aslin, 2004; Christiansen and Chater, 2008).

Acknowledgements

The ideas in this paper have benefited from exchanges with many people, including Adam Albright, Todd Bailey, Joe Pater, Katya Pertsova, Jason Riggle, Jennifer Smith, Anne-Michelle Tessier, and two anonymous reviewers, as well as audiences at the Phonologization Symposium at the University of Chicago and the West Coast Conference on Formal Linguistics at UCLA in 2008. A special debt is owed to Chris Wiesen for expert statistical advice, and to Abby Spears Wright for assistance in running participants and transcribing voice responses. Portions of this work were funded by the College of Arts and Sciences of the University of North Carolina. Remaining errors are the author's fault alone. Email may be addressed to `moreton@unc.edu`. This paper was first submitted on July 17, 2011. The current version is as of December 29, 2011. To appear (with minor changes) in the *Journal of Memory and Language*.

Appendix: Analysis of the perceptron learner

This appendix derives analytic expressions for the relation between performance and the number of training trials for the perceptron learner and training regime discussed in Section 7.3. We consider first the simpler HV condition, then the HH condition.

In the HV condition, there are five input units. The four stimuli produce the patterns of input activation shown in Table 20. Learning is error-driven. The update rule is $\mathbf{w} \leftarrow \mathbf{w} + \eta(t - o)\mathbf{i}$, where t and o were the correct and actual outputs on that trial ($+1$ or -1), $\mathbf{i}$ the vector of input-unit activations (each $+1$ or -1), and η is the learning rate, which is assumed to be very small.

When an error occurs on a positive stimulus (i.e., when $(+,+)$ or $(-,-)$ evokes the output -1), then $t - o = 2$. The two positive stimuli are distinguished from each other only by their labels; the network and training regime treat them exactly alike. Hence the expected change in $\mathbf{w}$ after an error on a positive trial is

Table 20: Input activation patterns produced by the four stimuli when applied to the network of Figure 1. Unit 5, $\alpha\alpha$, is omitted in the HV condition, present in the HH condition.

	Unit					
	0	1	2	3	4	5
Stimulus	"on"	$++$	$+-$	$-+$	$--$	$\alpha\alpha$
$(+,+)$	1	1	-1	-1	-1	1
$(+,-)$	1	-1	1	-1	-1	-1
$(-,+)$	1	-1	-1	1	-1	-1
$(-,-)$	1	-1	-1	-1	-1	1

$$E[\Delta\mathbf{w} \mid t = 1 \neq o] = 2\eta\frac{1}{2}\left((1,1,-1,-1,-1) + (1,-1,-1,-1,1)\right) \tag{1}$$

$$= \eta(2,0,-2,-2,0) \tag{2}$$

Likewise, when a negative stimulus is mistakenly classified as positive, $t-o = -2$, and

$$E[\Delta\mathbf{w} \mid t = -1 \neq o] = -2\eta\frac{1}{2}\left((1,-1,1,-1,-1) + (1,-1,-1,1,-1)\right) \tag{3}$$

$$= \eta(-2,2,0,0,2) \tag{4}$$

By symmetry of the net and the training regime, both kinds of errors are always equally probable, so the expected change to $\mathbf{w}$ when an error occurs is

$$E[\Delta\mathbf{w} \mid t \neq o] = \frac{1}{2}\left(\eta(2,0,-2,-2,0) + \eta(-2,2,0,0,2)\right) \tag{5}$$

$$= \eta(0,1,-1,-1,1) \tag{6}$$

i.e., averaged over many trials, w_1 and w_4 are incremented exactly as much as w_2 and w_3 are decremented, while increments and decrements to w_0 cancel out. Since all weights are initially zero, it follows that at any subsequent time, there is a w such that $\mathbf{w} = w(0,1,-1,-1,1)$.

By symmetry, the probability that a stimulus will be erroneously classified does not depend on the stimulus. Hence the probability of making an error on any stimulus is equal to the probability of making an error on $(+,+)$:

$$\begin{aligned}\Pr(\text{error}) &= \Pr(X \geq \mathbf{w} \cdot \mathbf{i}_{++}) &(7)\\ &= \Pr(X \geq w(0,1,-1,-1,1) \cdot (1,1,-1,-1,-1) &(8)\\ &= \Pr(X \geq 2w) &(9)\end{aligned}$$

where X is the noise source, a logistically-distributed random variable with mean 0 and variance 1. By a property of the logistic distribution,

$$\begin{aligned}\Pr(X \geq 2w) &= \left(1 - \frac{1}{1+e^{-2w/s}}\right) &(10)\\ &= \frac{1}{e^{2w/s}+1} &(11)\end{aligned}$$

where $s = \sqrt{3}/\pi$ (Balakrishnan and Nevzorov, 2003, 198). Passing to a continuous approximation, we have

$$\frac{dw}{dt} = \eta \frac{1}{e^{2w/s}+1} \qquad (12)$$

Integrating on both sides, we have

$$t = \frac{1}{\eta}\left(w + \frac{1}{2}se^{2w/s} + c\right) \qquad (13)$$

The boundary condition $w = 0$ at $t = 0$ yields $c = -s/2$. This equation relates training to weights, but our actual aim is to relate it to performance. It will be convenient to express performance in terms of the odds ratio $\Omega = \Pr(\text{correct})/\Pr(\text{incorrect})$. From Equation 11 we have $\Omega = e^{2w/s}$, and so $w = (s/2)\ln\Omega$. Hence Equation 13 becomes

$$t = \frac{s}{2\eta}(\Omega + \ln(\Omega) - 1) \qquad (14)$$

The dotted curve in Figure 2 was obtained by converting proportion correct (the vertical axis) to odds and then using Equation 14 to find the number of trials corresponding to that level of performance.

A similar analysis is applicable to the HH learner. The extra input unit does not break the symmetry, since it emits as much positive activation to the positive stimuli as it does negative activation to the negative ones. The weight proportions are given by $\mathbf{w} = w(0, 1, -1, -1, 1, 2)$. The probability of an error is now

$$\begin{aligned} \Pr(\text{error}) &= \Pr(X \geq \mathbf{w} \cdot \mathbf{i}_{++}) \quad (15) \\ &= \Pr(X \geq w(0, 1, -1, -1, 1, 2) \cdot (1, 1, -1, -1, -1) \quad (16) \\ &= \Pr(X \geq 4w) \quad (17) \\ &= \frac{1}{e^{4w/s} + 1} \quad (18) \end{aligned}$$

and so the relation between training and odds is

$$t = \frac{s}{4\eta}(\Omega + \ln(\Omega) - 1) \quad (19)$$

i.e., the HH learner reaches any given criterion in half the time required by the HV learner. Equation 19 is the source of the solid curve in Figure 2.

References

Alderete, J. and S. A. Frisch (2008). Dissimilation in grammar and the lexicon. In P. de Lacy (Ed.), *The Cambridge handbook of phonology*, Chapter 16, pp. 379–398. Cambridge, England: Cambridge University Press.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review 98*, 409–429.

Archangeli, D. and D. Pulleyblank (1994). *Grounded Phonology.* Cambridge, Massachusetts: MIT Press.

Archangeli, D. and D. Pulleyblank (2011). Harmony. In P. de Lacy (Ed.), *The Cambridge handbook of phonology*, Chapter 15, pp. 353–378. Oxford, England: Cambridge University Press.

Arnold, P. G. and G. H. Bower (1972). Perceptual conditions affecting ease of association. *Journal of Experimental Psychology 93*(1), 176–180.

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

Baković, E. (2011). Local assimilation and constraint interaction. In P. de Lacy (Ed.), *The Cambridge handbook of phonology*, Chapter 14, pp. 335–352. Oxford, England: Cambridge University Press.

Balakrishnan, N. and V. B. Nevzorov (2003). *A primer on statistical distributions.* Hoboken: Wiley-IEEE.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review 94*(2), 114–147.

Boersma, P. (1997). Functional Optimality Theory. *Proceedings of the Institute of Phonetic Sciences of University of Amsterdam 21*, 37–42.

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

Boersma, P. and J. Pater (2008, May). Convergence properties of a gradual learning algoritm for Harmonic Grammar. MS.

Boersma, P. and D. Weenink (2010). PRAAT Version 5.1.31. Software, www.praat.org.

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

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

Bye, P. (2011). Dissimilation. In M. van Oostendorp, C. Ewen, and E. Hume (Eds.), *The Blackwell companion to phonology*, Chapter 63. Wiley-Blackwell.

Campbell, L. (2004). *Historical linguistics: an introduction* (2nd ed.). Cambridge, Massachusetts: MIT Press.

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology 27*(3), 345–392.

Chambers, K. E., K. H. Onishi, and C. Fisher (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition 87*, B69–B77.

Chambers, K. E., K. H. Onishi, and C. Fisher (2010). A vowel is a vowel: generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition 36*(3), 821–828.

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

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

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

Ciborowski, T. and D. Price-Williams (1974). The influence of rule structure and problem composition on conceptual learning among rural Hawaiian children. Technical Report 75, Kamehameha Early Education Program, Kamehamea Schools and Bernice P. Bishop Estate, 1850 Makuakane Street, Honolulu, Hawaii 96817.

Clements, G. N. (1985). The geometry of phonological features. In J. A. Goldsmith (Ed.), *Phonological theory: the essential readings*, pp. 201–223. Malden: Blackwell.

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

Creel, S. C., E. L. Newport, and R. N. Aslin (2004). Distant melodies: statistical learning of non-adjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition 30*, 1119–1130.

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

Cristià, A., A. Seidl, and A. Francis (2011). Phonological features in infancy. In G. N. Clements and R. Ridouane (Eds.), *Where do phonological contrasts come from? Cognitive, physical, and developmental bases of phonological features*, pp. 303–326. Amsterdam and Philadelphia: John Benjamins.

Dell, G. S., D. R. Adams, and A. S. Meyer (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*(6), 1355–1367.

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

Endress, A. D., G. Dehaene-Lambertz, and J. Mehler (2007). Perceptual constraints and the learnability of simple grammars. *Cognition 105*(2), 247–299.

Endress, A. D. and J. Mehler (2010). Perceptual constraints in phonotactic learning. *Journal of Experimental Psychology: Human Perception and Performance 36*(1), 235–250.

Endress, A. D., B. J. Scholl, and J. Mehler (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General 134*(3), 409–419.

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

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

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

Frisch, S. A. and B. Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language 77*, 91–106.

Gallistel, C. R., A. L. Brown, S. Carey, R. Gelman, and F. C. Keil (1991). Lessons from animal learning for the study of cognitive development. In S. G. Carey and R. G. Gelman (Eds.), *The epigenesis of mind*, Chapter 1, pp. 3–36. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language 51*, 586–603.

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

Gottwald, R. L. (1971a). Attribute-response correlations in concept attainment. *American Journal of Psychology 84*(3), 425–436.

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

Guion, S. G. (1996). *Velar palatalization: coarticulation, perception, and sound change*. Ph. D. thesis, University of Texas, Austin.

Halle, M. (1962). A descriptive convention for treating assimilation and dissimilation. *MIT Research Laboratory of Electronics Quarterly Progress Report 66*(XVIII), 295–296.

Hansson, G. Ó. (2008). Diachronic explanations of sound patterns. *Language and Linguistics Compass 2*, 859–893.

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

Hayes, B., K. Zuraw, P. Siptár, and Z. Londe (2008, September). Natural and unnatural constraints in Hungarian vowel harmony. MS, Department of Linguistics, University of California, Los Angeles.

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

Heinz, J., C. Rawal, and H. Tanner (2011). Tier-based Strictly Local languages for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon.

Hockett, C. F. (1960). The origin of speech. In W. S.-Y. Wang (Ed.), *Human communication: language and its psychobiological bases*, pp. 4–12. San Francisco: W. H. Freeman.

Hume, E. and K. Johnson (2001). A model of the interplay of speech perception and phonology. In E. Hume and K. Johnson (Eds.), *The role of speech perception in phonology*, pp. 3–26. San Diego: Academic Press.

Hunt, E. B. and C. L. Hovland (1960). Order of consideration of different types of concepts. *Journal of Experimental Psychology 59*(4), 220–225.

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

Itô, J. and R. A. Mester (1995). Japanese phonology. In J. Goldsmith (Ed.), *The handbook of phonological theory*, Chapter 29, pp. 817–838. Cambridge, Massachusetts: Blackwell.

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

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language 59*, 434–446.

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

Kapatsinski, V. (2011). Modularity in the channel: the link between separability of features and learnability of dependencies between them. In W. S. Lee and E. Zee (Eds.), *Proceedings of the XVIIth International Congress of Phonetic Sciences (ICPhS)*, pp. 1022–1025.

Kim, H. (2001). A phonetically based account of phonological stop assibilation. *Phonology 18*(1), 81–108.

King, W. L. (1966). Learning and utilization of conjunctive and disjunctive classification rules: a developmental study. *Journal of Experimental Child Psychology 4*(3), 217–231.

Kirchner, R. (1998). *An effort-based approach to consonant lenition.* Ph. D. thesis, University of California, Los Angeles.

Köhler, W. (1941). On the nature of associations. *Proceedings of the American Philosophical Society 84*, 489–502.

Koo, H. and J. S. Cole (2006). On learnability and naturalness as constraints on phonological grammar. In *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, Athens, Greece, August 28-30*.

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

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

Laughlin, P. R. (1968). Focusing strategy for eight concept rules. *Journal of Experimental Psychology 77*(4), 661–669.

Laughlin, P. R. and R. M. Jordan (1967). Selection strategies in conjunctive, disjunctive, and biconditional concept attainment. *Journal of Experimental Psychology 75*(2), 188–193.

Leben, W. R. (1973). *Suprasegmental phonology.* Ph. D. thesis, Massachusetts Institute of Technology.

Lee, S.-S. (1981). The role of memory aids in two instructiona paradigms: learning and utilization of logical rules. *Instructional Science 10*, 123–133.

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

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

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

Magri, G. (2008). Linear methods in Optimality Theory: a convergent incremental algorithm that performs both promotion and demotion. MS, Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Majerus, S., L. Mulder, T. Meulmans, and F. Peters (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language 51*, 297–306.

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

Marler, P. (1991). The instinct to learn. In S. Carey and R. Gelman (Eds.), *The epigenesis of mind*, Chapter 2, pp. 37–66. Hillsdale: Erlbaum.

McCarthy, J. J. (1986). OCP effects: gemination and antigemination. *Linguistic Inquiry 17*(2), 71–100.

McCarthy, J. J. (1988). Feature Geometry and dependency: a review. *Phonetica 45*, 84–108.

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

Minsky, M. and S. Papert (1969). *Perceptrons: an introduction to computational geometry.* Cambridge, Massachusetts: MIT Press.

Mitchell, T. M. (1980 [1990]). The need for biases in learning generalizations. In J. W. Shavlik and T. G. Detterich (Eds.), *Readings in machine learning*, Chapter 2.4.1, pp. 184–191. San Mateo, California: Morgan Kaufman.

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

Moreton, E. (2010). Underphonologization and modularity bias. In S. Parker (Ed.), *Phonological argumentation: essays on evidence and motivation*, Chapter 3, pp. 79–101. London: Equinox.

Moreton, E. and J. Pater (to appear). Learning artificial phonology: a review. *Language and Linguistics Compass*.

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

Nevins, A. (2010). Two case studies in phonological universals: a view from artificial grammars. *Biolinguistics 4*(2), 218–233.

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

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

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

Ohala, J. J. and D. Feder (1994). Listeners' normalization of vowel quality is influenced by 'restored' consonantal context. *Phonetica 51*, 111–118.

Onishi, K. H., K. E. Chambers, and C. Fisher (2002). Learning phonotactic constraints from brief auditory experience. *Cognition 83*, B13–B23.

Pater, J. (2008). Gradual learning and convergence. *Linguistic Inquiry 39*(2), 334–345.

Pater, J. and A.-M. Tessier (2006). L1 phonotactic knowledge and the L2 acquisition of alternations. In R. Slabakova, S. Motrul, P. Prévost, and L. White (Eds.), *Inquiries in linguistic development: in honor of Lydia White*, pp. 115–131. Benjamins.

Peperkamp, S., K. Skoruppa, and E. Dupoux (2006). The role of phonetic naturalness in phonological rule acquisition. In D. Bamman, T. Magnitskaia, and C. Zoller (Eds.), *Papers from the 30th Boston University Conference on Language Development (BUCLD 30)*, pp. 464–475.

Pinker, S. (1979). Formal models of language learning. *Cognition 7*, 217–283.

Pons, F. and J. M. Toro (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition 116*, 361–367.

Prentice, W. C. H. and S. E. Asch (1958). Paired associations with related and unrelated pairs of nonsense-figures. *American Journal of Psychology 71*, 247–254.

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

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

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

Rescorla, R. A. (1986). Two perceptual variables in within-event learning. *Animal Learning and Behavior 14*(4), 387–392.

Rescorla, R. A. (2008). Evaluating conditioning of related and unrelated stimuli using a compound test. *Learning and Behavior 36*(2), 67–74.

Rescorla, R. A. and D. J. Gillan (1980). An analysis of the facilitative effect of similarity on second-order conditioning. *Journal of Experimental Psychology: Animal Behavior Processes 6*(4), 339–351.

Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In

A. H. Black and W. F. Prokasy (Eds.), *Classical conditioning*, Volume II: Current research and theory. New York: Appleton–Century–Crofts.

Rogers, C. J. and P. J. Johnson (1973). Attribute identification in children as a function of stimulus dimensionality. *Journal of Experimental Child Psychology 15*(2), 216–221.

Rose, S. and R. Walker (2004). A typology of consonant agreement as correspondence. *Language 80*(3), 475–531.

Rose, S. and R. Walker (2011). Harmony systems. In J. Goldsmith, J. Riggle, and A. C. L. Yu (Eds.), *The handbook of phonological theory* (2nd ed.)., Chapter 8, pp. 240–290. Malden, Massachusetts: Blackwell.

Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington, D.C.: Spartan Books.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language 47*, 172–196.

Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science 12*(4), 110–114.

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

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

Seidl, A. and E. Buckley (2005). On the learning of arbitrary phonological rules. *Language Learning and Development 1*(3 & 4), 289–316.

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

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

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

Snow, C. E. and M. S. Rabinovitch (1969). Conjunctive and disjunctive thinking in children. *Journal of Experimental Child Psychology 7*(1), 1–9.

Sproat, R. and O. Fujimura (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics 21*, 291–311.

Toro, J. M., M. Nespor, J. Mehler, and L. L. Bonatti (2008). Finding words and rules in a speech stream. *Psychological Science 19*(2), 137–144.

Toro, J. M., M. Shukla, M. Nespor, and A. D. Endress (2008). The quest for generalizations over consonants: asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception and Psychophysics 70*(8), 1515–1525.

Vance, T. J. (1987). *An introduction to Japanese phonology.* Albany, New York: State University of New York Press.

Walsh Dickey, L. (1997). *The phonology of liquids.* Ph. D. thesis, University of Massachusetts, Amherst, Massachusetts.

Warker, J. A. and G. S. Dell (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition 32*, 387–398.

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

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

Yeshurun, Y., M. Carrasco, and L. T. Maloney (2008). Bias and sensitivity in two-interval forced choice procedures: tests of the difference model. *Vision research 48*, 1837–1851.