Probabilistic Feature Attention as an alternative to variables in phonotactic learning

Brandon Prickett – blprickett@outlook.com

Abstract

Since Halle (1962), explicit algebraic variables (often called *alpha notation*) have been commonplace in phonological theory. However, Hayes and Wilson (2008) proposed a variable-free model of phonotactic learning, sparking a debate about whether such algebraic representations are necessary to capture human phonological acquisition. While past experimental work has found evidence that suggested a need for variables in models of phonology (Berent et al. 2012, Moreton 2012, Gallagher 2013), this paper presents a novel mechanism, *Probabilistic Feature Attention* (PFA), that allows a variable-free model of phonotactics to predict a number of these phenomena. Additionally, experimental results involving phonological generalization that cannot be explained by variables are captured by this novel approach. These results cast doubt on whether variables are necessary to capture human-like phonotactic learning and provide a useful alternative to such representations.

1. Introduction

Halle (1962) first proposed using explicit, algebraic variables (often called *alpha notation*) in phonological representations to better describe assimilatory and dissimilatory patterns. As an example, he showed that the representation for a voicing assimilation process could be simplified by using variables in place of more standard feature values like [+/−]. This is demonstrated below in (1):

(1) *Voicing Assimilation with Variables*

[−Sonorant] → [αVoice] / _[αVoice]

The example above shows a rule in which all obstruents must agree in voicing with the following segment. This agreement is represented by using the variable α as the value for the voicing features both in the output of the rule and the rule's context. This denotes that the value for that feature in those two segments can be *either* [+] *or* [−], but that it must be the same across the two sounds. Without the use of this algebraic representation, Halle (1962) notes that two rules would be required to represent the same process: [−Sonorant] → [+Voice] / _[+Voice] and [−Sonorant] → [−Voice] / _[−Voice], each handling one of the relevant environments that voicing assimilation can take place in.

Structures like the variables Halle (1962) proposed are the standard way of representing assimilatory and dissimilatory patterns in phonological theory, in both rule-based and constraint-based frameworks. These include autosegmental spreading rules (J. Goldsmith 1976), surface correspondence constraints (Rose & Walker 2004), and SHARE constraints (McCarthy 2010).[1] Whether such mechanisms are necessary for a model of phonological learning has primarily been explored using connectionist models of language (see, e.g. Gasser 1993, Marcus et al. 1999, Alhama & Zuidema 2018). However, Hayes and Wilson's (2008) variable-free maximum entropy phonotactic learner has made the debate over variables relevant to more standard phonological frameworks. Since their model was proposed, a number of experimental findings have been presented that suggest variables *are* necessary for capturing the kind of phonological learning and generalization observed in humans (Berent et al. 2012, Moreton 2012, Gallagher 2013).

In this paper, I use a novel training technique for variable-free phonotactic models, termed *Probabilistic Feature Attention*, which introduces structured errors into a learner's representations. These errors cause different segments in the model's training data to sometimes become ambiguous with one another, with ambiguity being more likely between segments that share a larger number of feature values. I show that this allows the model to generalize and learn in ways that have previously been explained using algebraic variables (Berent 2013), suggesting that they may not be as necessary as past work suggested. Additionally, I show that Probabilistic Feature Attention accounts for a phenomenon involving phonotactic generalization that variables cannot explain: *Similarity-based Generalization* (Cristia et al. 2013).

The rest of the paper will be organized as follows: §2 will give further background on phonotactic learning and variables, §3 will describe Probabilistic Feature Attention, §4 will describe four past experiments and simulations of those experiments using a model equipped with Probabilistic Feature Attention, and §5 will discuss.

[1] Note that these representations are equivalent to a specific implementation of variables. See Prickett (2020) for more on the differences between them and the kind of variables Halle (1962) first proposed.

2. Background

2.1.Variables in theories of phonology

Throughout this paper, the word *variable* will be used to describe any kind of representation that creates a dependency between two different feature values (Halle 1962, Marcus 2001, Berent 2013). For example, the constraint "*[αVoice][αVoice]" is violated by any sequence that has two adjacent segments that agree in voicing, because the *α* represents an identity relationship between the value of each segment's voicing feature.

While classic rule-based theories commonly made use of such variables (Halle 1962, Chomsky & Halle 1968) or methods that were similar (e.g. J. Goldsmith 1976), early connectionist models of phonology (see, e.g. Hare 1990, Corina 1991, Gasser & Lee 1992) lacked any kind of explicit symbolic representations (although, see Smolensky & Legendre 2006). However, when modeling reduplication with a recurrent neural network, Gasser (1993:6) suggested that such models would need "a variable of a sort" to learn the process in a human-like way.

Marcus et al. (1999) explored this idea by training both human infants and a variable-free neural network on reduplicative patterns and comparing their generalization. Specifically, they exposed seven-month-old infants to artificial languages whose words followed one of three surface patterns: ABA, ABB, or AAB, where each letter represents a syllable and repeated letters represent identical syllables. For example, in the AAB language, the first two syllables in a word would always be identical to one another (e.g. [wowofe]), similar to the reduplicated forms seen in many natural languages (Štekauer, Valera, & Körtvélyessy 2012).

The infants were then tested on words made up of novel segments that either conformed to or violated the pattern they learned in training. Infants listened to conforming words for significantly shorter periods of time than nonconforming words (see Rabagliati, Ferguson, & Lew-Williams 2019 for evidence of the reliability of these findings). Marcus et al. (1999) took this to mean that humans could generalize this pattern to words made of novel segments. However, when trained on the same items, a variable-free neural network was unable to generalize the pattern to new sounds, which they took as evidence of algebraic representations being necessary to capture this behavior.

Patterns other than reduplication have also been used as evidence that theories of phonology should make use of explicit variables. For example, in Hebrew, the first two consonants in a root

are not allowed to be identical (Greenberg 1950, Berent & Shimron 1997). Berent (2013) summarized a number of experiments meant to test whether Hebrew speakers use an algebraic representation when learning this restriction. The first experiment (Berent & Shimron 1997) asked Hebrew speakers to judge novel words that were made using native Hebrew phonemes. Novel words that violated the restriction were given significantly lower acceptability judgments than novel words that did not. This showed that the participants were not simply memorizing which words belonged to their language and were extracting some sort of general phonotactic pattern from their lexicon.

The next experiment (Berent et al. 2002) tested whether Hebrew speakers generalized the consonant restriction to novel phonemes. The sounds [dʒ], [tʃ], [w], and [θ], all of which are absent from the segment inventory of Hebrew, were used to create stimuli that either violated or conformed to the identity-based phonotactic pattern. Speakers again judged words whose first two consonants were identical as being worse in terms of grammaticality than those whose consonants differed. Based on these results, Berent (2013) suggested that algebraic variables must be present in the phonological grammars of native Hebrew speakers.

The generalization that Berent and colleagues observed in Hebrew and that Marcus et al. (1999) observed in artificial language learning will be called *Identity-based Generalization* for the remainder of this paper, since it involves generalizing a pattern of identity-based dependencies (e.g. "the first consonant in a root cannot be *identical* to the second consonant") to novel phonemes. Berent et al. (2012) explored whether the generalization performed by Hebrew speakers in the above-mentioned studies could be captured by a version of the variable-free maximum entropy phonotactic learner proposed by Hayes and Wilson (2008). They found that this was not the case and proposed a version of the Hayes and Wilson (2008) model that included explicit variables in its representations of phonological patterns.

To further explore whether the Hayes and Wilson (2008) model could capture Identity-Based Generalization, Gallagher (2013) compared the model's performance to that of humans in an artificial language learning experiment (see Linzen & Gallagher 2017, Tang & Baer-Henney 2019 for similar results). Specifically, in Gallagher's (2013) Experiment 2, participants were trained on a pattern that prohibited words with two voiced consonants, except when those consonants were identical. In training, the only identical pairs of consonants that participants

were exposed to were [b...b] and [d...d]. After training, participants were tested on withheld words like [g…g] and [d…g].

If participants generalized the identity-based pattern to the novel pair of identical consonants, [g...g], they would be more likely to treat the [g…g] words as exceptions to the restriction against two voiced consonants than a word like [d…g] that didn't have identical voiced segments. This is exactly what Gallagher (2013) found: participants were more likely to treat [g…g] items as grammatical than the [d…g] ones. Gallagher (2013) also found that the Hayes and Wilson (2008) model could not achieve the kind of Identity-Based Generalization she observed in her participants' responses without adding explicit variable-like representations to its architecture and interpreted this result as evidence for variables in the participants' phonological grammars.

However, generalization is not the only aspect of phonotactic learning that the presence of variables can affect—Gallagher's (2013) Experiment 1 demonstrated another behavior that models with variables can predict: *Identity Bias*. That is, participants trained to treat [b…b], [d…d], and [g…g] words as exceptions to the voicing restriction described above learned this pattern with higher accuracy than participants who were trained to treat an arbitrary set of words as exceptional. Gallagher (2013) showed that the Hayes and Wilson (2008) model also failed to capture this Identity Bias when it lacked any kind of variable-like mechanisms.

Moreton (2012) also observed a bias in phonotactic learning that he attributed to algebraic representations: *Intradimensional Bias*. In his experiments, subjects were either trained on phonotactic restrictions that involved multiple features across two sounds (e.g. [+High] vowels always occurring before [+Voice] consonants) or the same feature across sounds (e.g. [+High] vowels always occurring before other [+High] vowels). Participants learning patterns that involved a single feature across segments were more accurate in testing than those learning the two-feature pattern. Moreton (2012) showed that this bias could be replicated with a neural network (specifically, the *Configural Cue Model*; Gluck & Bower 1988) that was equipped with variables in its set of input features.[2] This was implemented with an input node labeled "[αHigh][αHigh]" that activated whenever the model was given inputs whose consonants matched in height, similar to the variable-like representations added to the Hayes and Wilson (2008) learner by Berent et al (2012) and Gallagher (2013).

[2] This bias has also been shown to emerge in recurrent neural networks without variables (Doucette 2017).

2.2.Variable-free models of phonotactic learning

This section describes Hayes and Wilson's (2008) original, variable-free model, as well as a related model (Pater & Moreton 2014, Moreton, Pater, & Pertsova 2017) that acts as the foundation for the phonotactic learner used in the current paper.

Hayes and Wilson (2008) proposed a maximum entropy phonotactic learner that uses constraints and constraint weights to store information about which surface forms are grammatical in a language. These constraints take the form "*X", "*XY", "*XYZ", etc., where X, Y, and Z stand for sounds or classes of sounds represented as a bundle of valued features. These feature bundles take the form [$\pm F_1 \ldots \pm F_i$], where *F* represents different features that are each valued as either [+] or [−]. An example of such a constraint would be "*[+Voice][−Labial, +Continuant]" which would match any bigram in which the first sound is [+Voice] and the second sound is both [−Labial] *and* [+Continuant].

Any word with a sequence of sounds that matches the description of a constraint would incur violations for that constraint equal to the number of times the illegal sequence occurs. The count of violations for a given word is multiplied by -1 and these negative violation counts are then multiplied by their constraint's weight. The sum of all these weighted violations is a word's *harmony*, which is then exponentiated and normalized to calculate a word's expected probability in whatever language the model has been trained on. This is demonstrated in Equations (1) and (2), where H_w is the harmony of word w, v_{cw} is the number of violations that w incurs for constraint c, C is the set of all constraints, and W is the set of all possible words up to a certain length.

$$H_w = \sum_{c \in C} w_c(-v_{cw}) \tag{1}$$

$$\Pr(w) = \frac{e^{H_w}}{\sum_{w' \in W} e^{H_{w'}}} \tag{2}$$

The constraints for the Hayes and Wilson (2008) model are induced from the training data by sampling from the space of possible constraints over the course of learning and introducing only those that help increase the model's overall estimate for the training data's probability. The model learns the optimal weights for these constraints using conjugate gradient descent, with the weight-learning and constraint-inducing alternating throughout training.

Since feature values are restricted to being either "+" or "−", constraints have no way of representing relationships between the feature values in different sounds. For example, if a language's phonotactics enforced voicing dissimilation (i.e. two adjacent consonants must have different values for [Voice]), the model would have to use two constraints to represent this: "*[+Voice][+Voice]" and "*[−Voice][−Voice]". Berent et al. (2012) and Gallagher (2013) both point out that the number of constraints that the Hayes and Wilson (2008) learner uses to learn such patterns could be reduced if explicit, algebraic variables were introduced into the learner's representations. For example, if a constraint of the form "*[αVoice][αVoice]" were included, the dissimilation pattern above could be represented with this constraint alone. This lack of algebraic variables in the model's constraints affects both the generalizations that the model makes, as well as the rate at which particular patterns are learned (Berent et al. 2012, Gallagher 2013).

Another maximum entropy phonotactic learner, *GMECCS* ("Gradual Maximum Entropy with a Conjunctive Constraint Schema"; Pater and Moreton 2014, Moreton et al. 2017), behaves similarly to the Hayes and Wilson (2008) model and will be used for the simulations presented in this paper.[3] The equations for harmony and word probability are identical in GMECCS, and constraint weights are also optimized using a similar algorithm (standard gradient descent). However, the constraint set that the model uses was inspired by the Configural Cue Model (Gluck & Bower 1988), and includes every possible conjunction of the features needed to describe the relevant data. These constraints are all present throughout the learning process, rather than being induced.[4] This is illustrated in Table 1 for a scenario where the only possible features are [Voice] and [Continuant] and words are limited to being a single segment long.

	***[+v]**	***[-v]**	***[+c]**	***[-c]**	***[+v, +c]**	***[+v, -c]**	***[-v, +c]**	***[-v, -c]**
d	1			1		1		
z	1		1		1			
t		1		1				1
s		1	1				1	

Table 1. Violations for four possible words when the feature space is restricted to [Voice] and [Continuant], and words are limited to being one phoneme long. To save space, [Voice] has been shortened to [v] and [Continuant] has been shortened to [c].

[3] Two differences exist between the implementation of GMECCS used here and the original version of the model: (1) my version of GMECCS uses online learning, rather than batch learning, and (2) Moreton et al (2017) used constraints like "*[+Voice]$_{C2}$" which would only be violated by a word with a [+Voice] sound as its second consonant. Instead of these, I used more standard unigram constraints like those shown in columns 2-5 of Table 1 that apply to all of the segments in a word.

[4] Another difference between GMECCS and the Hayes and Wilson (2008) model is that the latter restricts constraint weights to being positive. The weights in GMECCS can be any real-valued number.

This table demonstrates a number of facts about the conjunctive constraint set used by GMECCS. First of all, the total number of constraints increases rapidly as the number of features and length of words increases. With only two features and words of length one, there are already eight constraints. If the number of features was increased to three, the number of total constraints would grow to 26. If the possible length of words was then increased to two, the number of constraints would be 702. This means that for practical purposes, any given simulation using a constraint set like GMECCS's must use the smallest feature space and word length necessary to capture the relevant training and testing data.

The table also shows the violations that each of the possible data points would have in this simplified scenario. Because of the exhaustive nature of the constraint set that GMECCS uses, each datum has a unique set of constraint violations. This, combined with a lack of any regularization priors, means that the model can perfectly match any probability distribution that it is trained on.

Like the Hayes and Wilson (2008) model, GMECCS lacks any explicit, algebraic variables. This means that it should have the same limitations regarding Identity-based Generalization, Identity Bias, and Intradimensional Bias as other variable free models. In §4, results are shown for GMECCS in a number of simulations that confirm this—without any added mechanisms, GMECCS fails on the same kind of tests that Berent et al. (2012) and Gallagher (2013) show the Hayes and Wilson (2008) model failing on.

3. Probabilistic Feature Attention

This section will introduce a novel mechanism, *Probabilistic Feature Attention* (PFA), that is designed to test whether variables are necessary to capture human learning and generalization, as past work has suggested (Marcus et al. 1999, Berent et al. 2012, Moreton 2012, Gallagher 2013).

PFA is inspired by the use of dropout in deep learning (Srivastava et al. 2014), which has been shown to aid variable-free neural networks in generalizing identity-based patterns (Prickett, Traylor, & Pater 2018). PFA is designed to be used with the kind of maximum entropy phonotactic learners described in §2.2 and demonstrates how certain kinds of errors in the learning process can cause such models to behave in a more human-like way. These errors are meant to be an abstract way of representing the fact that language acquisition often involves imperfect data, due to issues like misperception (see, e.g., Bailey & Hahn 2005) and constraints

on memory (see, e.g., Gathercole & Adams 1993). The core assumptions of PFA are: (1) learners do not attend to every feature in the representation of every form they are presented with, (2) this lack of attention creates ambiguity in the learner's input, and (3) in the face of ambiguity, learners err on the side of assigning constraint violations to the form.

In PFA, features are either completely attended to or completely dropped out of a stimulus's representation throughout the learning process.[5] This is implemented by stepping through each feature on each weight update and randomly deciding whether that feature will be attended to for that iteration of learning, with no limit on the number of features that can be attended to or ignored. The probability that any given feature will be attended to is a hyperparameter for the model and was .25 for all of the simulations presented in this paper.[6] This means that the probability in any given simulation that all of the features in a stimulus were attended to was equal to $.25^N$, where N is the number of features used in that simulation.

Because the probability of attending to all features is so small, the model is regularly faced with ambiguity. For example, consider the simplified scenario discussed in §2.2 in which [Voice] and [Continuant] are the only possible features and words are limited to being a single segment long. Now assume that PFA is being used to learn a language in which some subset of these four words is grammatical, and that in the current weight update the feature [Voice] is the only one that's being attended to. This would mean that [t] and [s] are completely ambiguous with one another (as are [z] and [d]). Rather than having a four-way distinction between all of the possible surface forms, the model is only able to see whether a sound is [+Voice] or [−Voice].

Rather than attempting to estimate constraint violations for the ambiguous data (as in, e.g. Tesar 2004, Jarosz 2013), when PFA is used, an ambiguous form is given all of the violations for all of the forms that it *might be*.[7] So, a sound that is ambiguous between [t] and [s] would violate

[5] The idea that features are not all attended to equally is not new and can be seen in Nosofsky's (1986) exemplar classification model and a number of its successors. In that model, a mechanism called *Selective Feature Attention* was used to determine which features were most important to attend to in a non-linguistic classification task. Weights were learned for each feature, and at the end of learning, the weights represented which features were important for that pattern.

[6] While this value remains constant for all of the simulations presented here, it could be slowly increased over the course of learning to speed up acquisition. Another consequence of doing this is that once the probability of attending to features reaches 1, the model would be guaranteed to eventually find an optimal set of constraint weights (Moreton et al. 2017).

[7] Note that sounds are not ambiguous with segments that are completely absent from a model's segment inventory. For example, if a model was not given any [+Labial, +Velar] sounds in its input files, then [+Velar] sounds would not violate the constraint *[+Labial], even when [Labial] is not attended to, since the model would have independent evidence that [?Labial, +Velar] sounds should not incur such violations.

all of the constraints that either [t] *or* [s] violate. In the constraint set from Table 1, this would mean that the only constraints that are not violated by such a sound are "*[+Voice]", "*[+Voice, +Continuant]", and "*[+Voice, −Continuant]". The violation vectors for all the ambiguous sounds in this simple example, given the different options for which features are attended to, are shown in Table 2.

		***[+v]**	***[-v]**	***[+c]**	***[-c]**	***[+v, +c]**	***[+v, -c]**	***[-v, +c]**	***[-v, -c]**
[voice]	**[-v, ?c]**		1	1	1			1	1
	[+v, ?c]	1		1	1	1	1		
[cont.]	**[?v, -c]**	1	1		1		1		1
	[?v, +c]	1	1	1		1		1	
None	**[?v, ?c]**	1	1	1	1	1	1	1	1

Table 2. Violations for all possible ambiguous words when the feature space is restricted to [Voice] and [Continuant], words are limited to being one phoneme long, and PFA is being used. The left-most column shows which features are being attended (if both [voice] and [cont.] were attended to, no ambiguity would be present). For example, [+Voice, ?Cont] is any voiced sound when the feature [Continuant] is not being attended to. To save space, [Voice] has been shortened to [v] and [Continuant] has been shortened to [c] and [cont.].

While PFA could hypothetically be paired with a number of models and learning algorithms, the results presented in §4 will all use GMECCS and stochastic gradient descent (specifically, a version of stochastic gradient descent in which one datum is presented to the learner at each weight update, sometimes referred to as *online learning*).[8] Stochastic gradient descent is a gradual, error-based algorithm that updates weight values based on how much each constraint contributed to either a correct or incorrect probability estimate for a particular datum. Following Moreton et al. (2017) and Hayes and Wilson (2008), the model is given access to the probability distribution over all possible data. That is, the model is trained to map attested forms to the probability with which they occur in the training data and unattested forms to a probability of 0, by being exposed to both kinds of mappings over the course of acquisition.[9]

[8] All of the results presented in this paper were also found with standard gradient descent (i.e. training in batch, where all of the data is presented at each weight update), which was the algorithm that GMECCS was originally paired with.

[9] While this could be considered a form of negative evidence for the model (since it is exposed to attested and unattested forms), exposure to forms that map to zero is necessary to find the appropriate weight updates when using gradient descent (see Hayes and Wilson:§3.3.2 for more discussion on this).

In PFA, ambiguity means that many of the model's probability estimates during training are made using ambiguous data. This causes updates to often happen in ways that do not necessarily improve the model's overall acquisition of a pattern. For example, consider the simple scenario from Tables 1 and 2, and a language in which only the sounds [s] and [t] are grammatical. As GMECCS learns this pattern, it will gradually assign more weight to constraints like "*[+Voice]" and less weight to constraints like "*[−Voice]". However, in any iteration in which [Voice] is not attended to, the model will be unable to learn anything about the restriction present in the language, since sounds will all either be [?Voice, +Continuant] or [?Voice, −Continuant]. These sounds would be ambiguous between the ones that are allowed in the language ([s] and [t]) and the ones that are banned ([z] and [d]).

In more complex patterns, this ambiguity can push the model in directions that causes it to represent the training data more poorly. For example, if a language only allowed [s] and [d], more constraints would be necessary to represent it. The model would need to add more weight to constraints like "*[+Voice, +Continuant]" and "*[−Voice, −Continuant]" (i.e. those that are violated by [z] and [t] respectively), while assigning less weight to constraints like "*[−Voice, +Continuant]" and "*[+Voice, −Continuant]" (i.e. those that are violated by [s] and [d], respectively). Every time the model sees data like [t] mapping to a probability of 0 or [d] mapping to a nonzero probability, it will correctly move the weights in these directions, as illustrated in the first two rows of Table 3. However, if it is presented with a [?Voice, −Continuant] segment mapping to a high probability (i.e. [d] with ambiguous voicing, as shown in the last row of Table 3), it will take weight away from both "*[−Voice, −Continuant]" *and* "*[+Voice, −Continuant]", since this ambiguous word violates both constraints. This would move probability onto both [d] and [t], even though [t] is ungrammatical in the language.

Actual Datum	**Features Attended**	**Datum seen by Model**	**Example Weight Updates**	**Model's Prob. Estimates**
pr(t)=0	[Vce.], [Cont.]	pr(t)=0	*[−Vce., −Cont.] ↑	pr(t) ↓
pr(d)=.25	[Vce.], [Cont.]	pr(d)=.25	*[+Vce., −Cont.] ↓	pr(d) ↑
pr(d)=.25	[Cont.]	pr([?Vce, −Cont.])=.25	*[−Vce., −Cont.] ↓ *[+Vce., −Cont.] ↓	pr(t) ↑ pr(d) ↑

Table 3. Illustration of PFA in which each row represents a separate weight update. The features [Voice] and [Continuant] are shortened to [Vce.] and [Cont.], respectively.

When PFA moves weights in a seemingly incorrect direction like this, over time it causes segments and sequences of segments that are featurally similar to be treated similarly by the model. This is because the fewer features two sounds differ in, the more likely they are to become ambiguous with one another due to PFA. Since variables' purpose in phonological representations is usually to capture inter-segmental, feature-based similarity (Halle 1962) and since PFA causes ambiguity to correlate with this same kind of similarity metric, the two mechanisms end up predicting qualitatively similar predictions for a number of tasks. In §4, I'll demonstrate this for three of these tasks that have been used to argue to variables in the literature.

4. Experiment simulations

This section will summarize four past experiments demonstrating various human behaviors in an artificial language learning context and show that a maximum entropy phonotactic learner with PFA can capture each experiment's results.[10] The first three experiments (§4.1-4.3) have all been used to argue for the existence of variables in the phonological grammar, while the fourth experiment (§4.4) *cannot* be accounted for using algebraic representations.

4.1. Identity-based Generalization

While Identity-based Generalization has been demonstrated in a number of contexts (e.g. Marcus et al. 1999, Berent et al. 2002), here I'll use Gallagher's (2013) Experiment 2 as a test case for PFA. In that experiment, participants were trained on a language in which surface forms were not allowed to contain two voiced consonants unless those consonants were identical to one another. Crucially, words containing the consonant pairs [d...g] and [g...g] were withheld from training to see if participants generalized the identity-based pattern of exceptionality to a phoneme they hadn't seen in this context—something that Gallagher (2013) showed the Hayes and Wilson (2008) model could not do.

Since the vowels in Gallagher's (2013) training data were irrelevant to the pattern at hand, I represented words in my simulations as a bigram of consonants and only consonantal features were used to create the model's set of constraints. The phonemes used in these simulations, as well as their corresponding feature values, are shown in Table 4.

[10] The code and training data for all simulations can be viewed at *[URL removed for anonymity]*.

	[voice]	[labial]	[dorsal]
b	+	+	-
p	-	+	-
d	+	-	-
t	-	-	-
g	+	-	+
k	-	-	+

Table 4. Features and phonemes used in simulating Gallagher's (2013) Experiment 2.

The training data consisted of every possible bigram of the relevant consonants, with output forms that Gallagher's (2013) participants were exposed to in training mapping to a probability of 0.143 (i.e. 1 divided by the number of attested items, 7) and all other forms mapping to a probability of 0. Each of these data are shown in Table 5.

Attested Forms (p=.143)	**Unattested Forms (p=0)**			
bb	bd	gb	kp	pt
bk	bg	gd	kt	tb
bt	bk	gg	pb	td
dd	bp	gk	pd	tg
dp	db	kb	pg	tk
gp	dg	kd	pk	tp
gt	dk	kg	pp	tt
	dt	kk		

Table 5. Training data used to simulate Gallagher's (2013) Experiment 2. Words that participants would have been exposed to in the experiment were mapped to a probability of .143 (1/7) in the simulation's training data, while words absent from the experiment exposure were mapped to a probability of 0.

First, the standard version of GMECCS (i.e. with no PFA) was trained on this data for 200 epochs,[11] with a learning rate of 0.05. At the end of this training, the model's constraint weights were used to estimate the probabilities of each of Gallagher's (2013) test items: [gg] and [dg] (both of which mapped to a probability of 0 in the model's training data). If the model behaved like her participants, it would predict a higher probability for [gg] since the form's identical consonants should allow it to act as an exception to the restriction on [+Voice][+Voice] sequences. The model's average estimates over 15 separate runs at the end of learning are shown in Figure 1 and are representative of the full course of acquisition. The probability that the model assigns to [dg] is always significantly higher than the probability it assigns to [gg], meaning that,

[11] Throughout this paper, I will use the word "epoch" to refer to a complete pass through the training data and "iteration" to refer to a single weight update of the model. Since all simulations consisted of multiple epochs, the model always saw each datum more than once; however, each time the datum was observed, the set of features that was attended to was resampled.

like the Hayes and Wilson (2008) model, the standard version of GMECCS is not able to capture Identity-based Generalization.

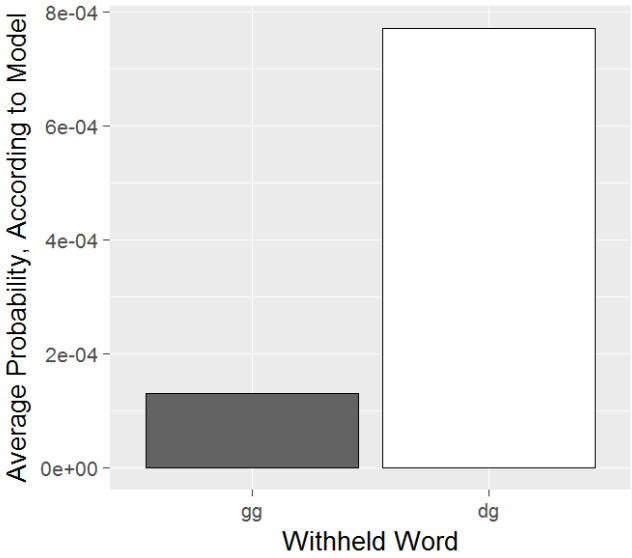

Figure 1. Results from GMECCS at the end of learning (200 epochs), after it was trained on the Gallagher (2013) Experiment 2 data with no variables and no PFA. Figure gives the average of 15 runs. No error bars are shown because the standard error of the mean across runs was too close to 0 to be visible in the figure.

This failure is a result of the fact that a conjunctive constraint set allows the model to keep track of unigram and bigram probabilities simultaneously, meaning that the [d] in the word [dg] will always allow it to have more probability than the word [gg]. This is illustrated in Table 6 with a subset of the constraints and data that were used in the simulation (and a simplified set of weights for those constraints to aid in interpretability).

	***[+vc, -dl]**	***[+vc, +dl]**	***[+vc, -dl][+vc, -dl]**	***[+vc, -dl][+vc, +dl]**	***[+vc, +dl] [+vc, +dl]**	**Pr.**
	weight=1	weight=1.2	weight=0.01	weight=2	weight=2	
dg	1	1		1		.09
gg		2			1	.08
dd	2		1			.83

Table 6. Illustration of why GMECCS fails to model Gallagher's (2013) results. Weights are simplified to be more interpretable. See Equations (1) and (2) for how the probabilities were calculated. For the sake of space, [Voice] is shortened to [vc] and [Dorsal] is shortened to [dl].

The weights in this table demonstrate a hypothetical, simplified scenario that illustrates why GMECCS fails to assign more probability to [gg] than [dg] throughout learning. Since both [dg] and [gg] are absent from the training data, the bigram constraints they each violate have equally high weights. Similarly, the bigram constraint that [dd] violates, "*[+Voice, −Dorsal][+Voice,

−Dorsal]", has a low weight because that form has a nonzero probability in the training data. If the model only used bigram constraints, it would assign an equal probability to [dd] and [gg].

However, the unigram constraints cause there to be a difference in the two withheld forms' predicted probabilities. The constraint "*[+Voice, −Dorsal]" drives this difference, since it's violated twice by the attested form [dd], meaning that this constraint is assigned a relatively low weight. However, "*[+Voice, +Dorsal]" receives a slightly higher weight since it's violated by [g] and this segment is less common in attested words in the training data. This asymmetry between [d] and [g] means that throughout learning, [dg] will receive a higher probability than [gg], even if the bigram constraints that they each violate have equal weights. This mirrors the results that Gallagher (2013) found with the Hayes and Wilson (2008) model—neither variable-free phonotactic learners could properly generalize identity-based exceptions.

Next, the version of GMECCS that uses PFA was trained on the same data, with the same hyperparameters. This model's estimations for the test data at the end of learning, averaged over 15 runs, are shown in Figure 2.[12] These results are also representative of the entire course of acquisition.

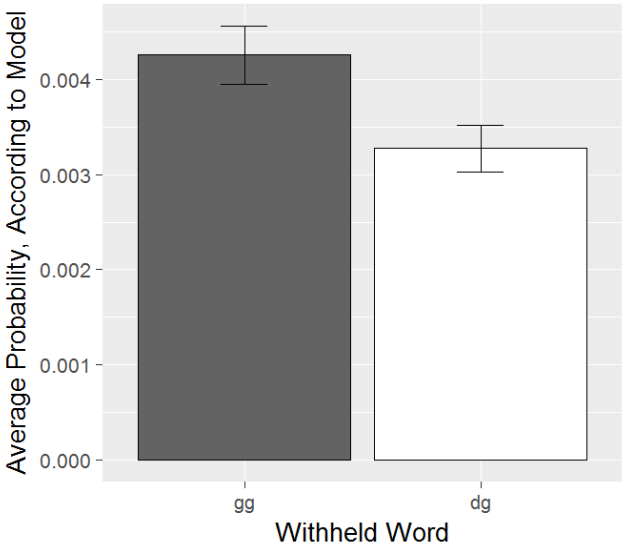

Figure 2. Results from GMECCS at the end of learning (200 epochs), after it was trained on the Gallagher (2013) Experiment 2 data with PFA. Figure gives the average of 15 runs. Error bars show the standard error of the mean.

[12] Note that the probabilities in Figures 1 and 2 are relatively low compared to the proportions of each form produced by humans in Gallagher's (2013) experiment. However, this is due to the fact that the model has to assign probabilities to all possible bigrams, while the results presented in Gallagher represent only the human responses that were either [g...g] or [d...g] words.

The model with PFA was able to capture the Identity-based Generalization observed in Gallagher's (2013) Experiment 2. [13] The word [gg] was consistently given higher probability than [dg]. This was a result of the fact that, over the course of learning, [dg] is more likely to become ambiguous with other unattested words than [gg] is. This can be demonstrated by focusing on the difference between the two test words: [dg] is a [−Dorsal][+Dorsal] sequence and [gg] is a [+Dorsal][+Dorsal] sequence. Table 7 shows all of the words that were mapped to a probability of 0 in the model's training data, organized by their values for the [Dorsal] feature in each segment.

Since there are more [−Dorsal][+Dorsal] words with a probability of 0 than their [+Dorsal][+Dorsal] counterparts, [dg] is more likely to become ambiguous with other zero-probability words in the training data. For example, any time the model sees unambiguous versions of [dg] or [gg] mapping to zero, it will assign higher weights to constraints that those words violate (as illustrated in the first two rows of Table 9).

[-dorsal][-dorsal]		**[+dorsal][-dorsal]**	**[-dorsal][+dorsal]**	**[+dorsal][+dorsal]**
tt	pd	kt	tk	kk
td	pp	kd	tg	kg
tp	pb	kp	dk	gk
tb	bd	kb	***dg***	***gg***
dt	bp	gb	pk	
db		gd	pg	
pt			bg	

Table 7. Words that had a probability of 0 in the simulations described above, organized by their phonemes' value for [dorsal]. The test words used by Gallagher (2013) are shown in bold italics.

However, if the model is presented with [dk] and is not attending to [Voice], the words [dk], [tk], [tg], and (crucially) [dg] will all be ambiguous with one another. Since [dk] maps to a probability of 0, and since all four of these data are ambiguous with one another when voicing is ignored, the learner will raise the weights of all the constraints that any of the four forms violate (as illustrated in the third row of Table 9). This will take probability away from [dg], even though the learner was not actually presented with that item during the relevant iteration.

[13] Thanks to an anonymous reviewer for pointing out that Gallagher's (2013) training data for this experiment lacked a set of words that would likely be present in a natural language with the same restriction: those with a [−voice]...[+voice] pair of consonants. To ensure this simplification didn't affect my results, I ran identical simulations that included the bigrams [tb], [kb], [pd], [pg], and [tg] mapping to nonzero probabilities in the training data. This simulation had similar results, except for the fact that the advantage that [gg] had over [dg] eventually disappeared over the course of learning.

Actual Datum	Features Attended	Datum seen by Model	Example Weight Updates	Model's Prob. Estimates
pr(dg)=0	[V], [D], [L]	pr(dg)=0	*[−D][+V,+D] ↑	pr(dg) ↓
pr(gg)=0	[V], [D], [L]	pr(gg)=0	*[+D][+V,+D] ↑	pr(gg) ↓
pr(dk)=0	[D], [L]	pr([?V, −L,−D][?V, −L,+D])=0	*[−D][−V,+D] ↑ *[−D][+V,+D] ↑	pr(dk) ↓ pr(dg) ↓

Table 8. Illustration of the PFA model learning the language from Gallagher's (2013) Experiment 2. Each row represents a separate weight update. The features [Voice], [Dorsal], and [Labial] are shortened to [V], [D], and [L], respectively.

This example would be representative of much of the model's learning process—as long as [Dorsal] is being attended to and at least one other feature is not, multiple items in the [−Dorsal][+Dorsal] column above will be ambiguous with [dg]. However, since there are only three other items in the [+Dorsal][+Dorsal] column, [gg] is much less likely to have its probability lowered in this way. This means that over the course of learning, PFA will cause the model to take more probability off of [dg] than [gg], mimicking an identity mapping, even though no actual identity relationship is being represented in the model's constraint set.

However, it's important to note that the success of the PFA model on this task was dependent on the feature set used in the simulation. For example, if a binary version of the feature [Coronal] had been included, the asymmetry demonstrated in Table 7 would disappear. In this case, the relevant difference between [dg] and [gg] is that the former's first segment is [+Coronal, −Dorsal] and the latter's is [−Coronal, +Dorsal]. There would then be an equal number of [+Coronal, −Dorsal][+Dorsal] and [−Coronal, +Dorsal][+Dorsal] words mapping to a probability of zero in the training data, meaning that the advantage for [gg] present in the previous results would be gone. To confirm this, I ran a simulation of Gallagher's (2013) experiment that was identical to the one described above, except for the fact that the feature [±Coronal] was included in the model's representations. The results from this are shown in Figure 3, in which the probability of [gg] is either lower than or not significantly different from the probability of [dg] for the entire course of learning.

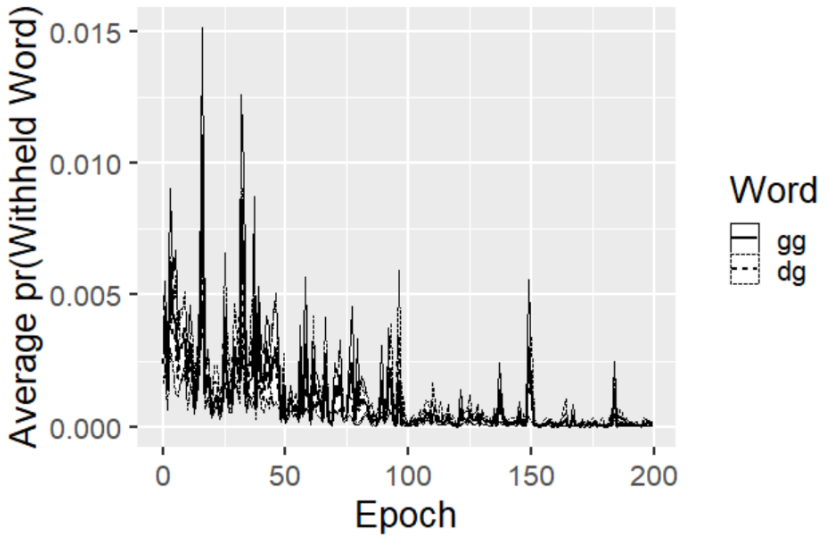

Figure 3. Results from a simulation of Gallagher's (2013) Experiment 2, in which [Coronal] is included in the model's representation.

This demonstrates that the model's predictions are dependent on the feature theory used, but this feature set is still not the most standard way of representing place of articulation—typically ternary features are used that can take on the values [−], [+], or unmarked (see, e.g. Hayes 2011). I leave testing the effects of an unmarked feature value on PFA to future work, since a number of nontrivial decisions must be made to implement such representations in this system.

To test how the model behaves with a feature theory that is more standard, and to see how generalizable the model's success at Identity-Based Generalization is, I ran a simulation that represents a much simpler scenario: a language with the segment inventory [i], [e], [o], and [u] that only allows words with identical vowels in them. This inventory can be divided up using the features [Back] and [High], both of which are binary in most standard feature theories. The training data for these simulations consisted of the words [ii], [ee], and [uu] mapping to a probably of .333 (1/3), and every other possible bigram using these vowels mapping to zero. The words [oo] and [ou] were used as test items and the model was asked to estimate their probability at each epoch of learning. If the model captured Identity-Based Generalization in this simulation, then it would assign higher probability to [oo] than to [ou], since the former has identical vowels. The results for this simulation at the 200th epoch (which was representative of the full course of acquisition) with the version of GMECCS that uses PFA are shown in Figure 4.

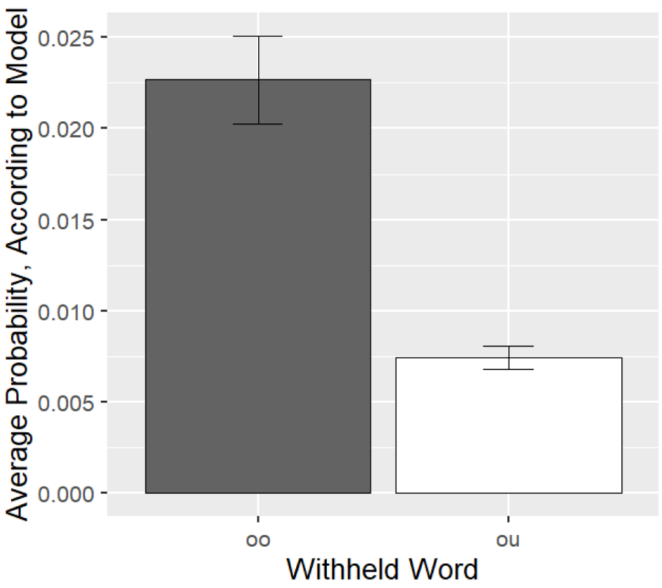

Figure 4. Results from the version of GMECCS that uses PFA. Training data represented a hypothetical, simplified scenario that tests the same kind of Identity-Based Generalization as Gallagher's (2013) Experiment 2. Test items were [oo] and [ou].

These results show that in a simplified scenario that uses a more standard set of features, the model with PFA can again capture Identity-Based Generalization. This is primarily due to two factors. First, as in the previous successful simulation, [ou] is more likely to become ambiguous with other unattested bigrams. This is because the relevant difference between [oo] and [ou] is that the former is [−High][−High] while the latter is [−High][+High]. The set of [−High][−High] bigrams mapping to a probability of zero is [oo], [oe], and [eo] while the set of [−High][+High] bigrams mapping to a probability of zero is [oi], [ou], [ei], and [eu]. This means that the test item [ou] is more likely to become ambiguous with other items that map to zero in training and will lose probability more quickly over the course of learning.

Furthermore, the test item [oo] will be more likely to become ambiguous with words that map to a probability that's *greater than zero* in the model's training data. This wasn't true in the simulations of Gallagher's (2013) experiment, due to the asymmetrical nature of the place of articulation features that I used. However, in this simulation, any time the feature [Back] is not attended to, the words [ee] and [oo] will be indistinguishable, and if this ambiguous datum is mapping to a probability of .333 (i.e. if it's actually [ee]), more probability will shift to *both* of these items. For [ou] to become ambiguous with an attested item, two features would need to be ignored, and any pair of features that would allow this to happen would also cause ambiguity between [oo] and the attested words.

This series of simulations reveals that while PFA *allows* a variable-free model to predict the kind of Identity-Based generalization that Gallagher (2013) observed, it only predicts this generalization when the feature theory and data set that the model is given allows at least one of two characteristics to be true: either identical withheld items are more similar to words that are present in a language's lexicon *or* withheld non-identical items are more similar to words that are absent from the lexicon.

4.2. Identity Bias

Instead of focusing on generalization, Gallagher's (2013) Experiment 1 demonstrated that the relative learning rate of different phonological patterns can be correctly predicted by models with explicit variables. The two patterns that Gallagher's (2013) Experiment 1 looked at were labeled the *Identity Language* and the *Arbitrary Language*. In the *Identity Language*, the pattern was identical to her Experiment 2 (described in §4.1), except there were no withheld word types in training (crucially, [g…g] words were now included). In the *Arbitrary Language*, there was still a restriction banning two voiced consonants in the same word, however, the exceptions to this restriction no longer had a systematic relationship to one another. That is, [b…d], [d…g], and [g…b] words were all allowed, and all other [+Voice]…[+Voice] words were ungrammatical.

For the simulations presented here, the words were again simplified to the consonants that made them up, and the same phonemes and features described in §4.1 were used. In the Identity language, all words that only had one voiced consonant, as well as those that had identical voiced consonants mapped to a nonzero probability of .111 (i.e., 1/9) and words that had two, non-identical voiced consonants mapped to a probability of zero. In the Arbitrary Language, the forms [bd], [dg], [gb], as well as all forms that lacked two voiced consonants mapped to a probability of .111 and every other bigram mapped to 0.

First, the standard version of GMECCS was trained separately on these two languages for 200 epochs each, with a learning rate of 0.05. The learning curves for these simulations, averaged over 30 runs in each language, are shown in Figure 5. If the model captured the humans' behavior, it would assign probability to grammatical words more quickly in the Identity Language than in the Arbitrary Language.

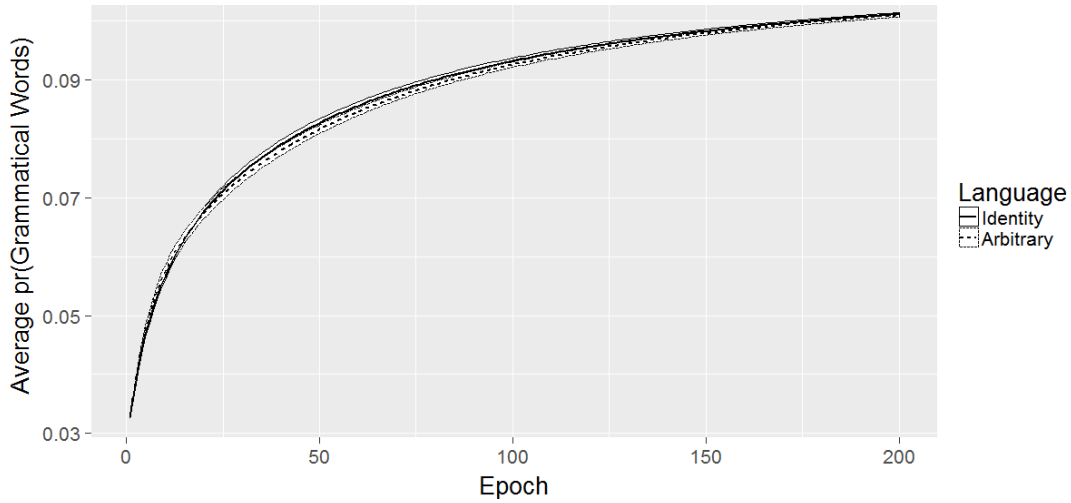

Figure 5. Learning curve for the model, when trained on the Gallagher (2013) Experiment 1 data with a version of GMECCS that has no variables and no PFA. Lines show the average of 30 runs. White space behind each line shows the standard error of the mean at that epoch.

While a small difference does seem to exist between the learning curves for the Identity and Arbitrary Languages, there is a large degree of overlap in their variance. The reason why GMECCS does not show a reliable bias for the Identity pattern is because it takes a similar number of constraints to describe both of the relevant languages (Berent et al. 2012, Gallagher 2013). Since the model's constraints have no way of representing inter-segmental similarity, as far as the model is concerned, the exceptional words in the Identity Language are just as arbitrary as those in its counterpart.

Figure 6 illustrates results from simulations with the version of GMECCS that uses PFA (averaged over 30 runs each, same hyperparameter values as above). This shows that with PFA, a consistent preference for the Identity Language is present in the learning curves past the 50th epoch. The Identity Language is easier for the learner because the datapoints that map to nonzero probabilities in that language are more likely to become ambiguous with one another in ways that aid the acquisition of the overall pattern. For example, [bb] and [dd] are ambiguous with one another at any point in learning where [Labial] is not being attended to. This means that they would both violate "*[+Voice, +Labial][+Voice, +Labial]" and "*[+Voice, −Labial][+Voice, −Labial]", so that these constraints would lose weight over the course of learning more quickly than if no ambiguity was present. Since assigning low weights to both of these constraints is the correct action to take while learning the pattern, this causes the Identity Language to be acquired relatively quickly.

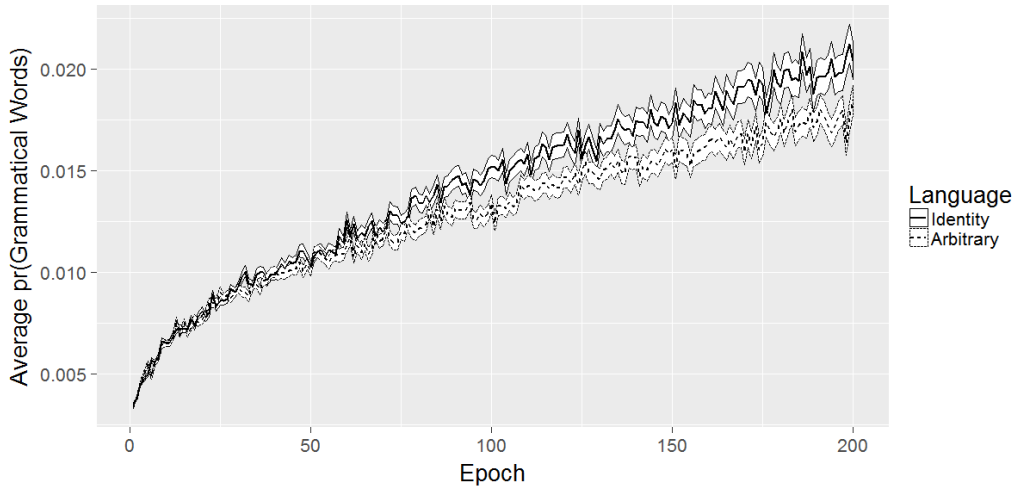

Figure 6. Learning curve for the model, when trained on the Gallagher (2013) Experiment 1 data with a version of GMECCS that uses PFA. Figure gives the average of 30 runs. White space behind each line shows the standard error of the mean at that epoch.

Since the Arbitrary Language's attested words do not have this kind of systematic similarity across data, the ambiguous words created from them do not end up violating constraints that assist the learning of the pattern as a whole. Instead, the random ambiguity just creates noise in the learning process for the Arbitrary Language, making it more difficult to acquire than Identity. This means that PFA can simulate a preference for learning identity-based patterns without using explicit, algebraic variables. Furthermore, unlike Identity-Based Generalization, this characteristic of PFA is not particularly sensitive to the data set and feature theory that's being given to the model.

4.3. Intradimensional Bias

Another learning bias that has been attributed to variables is Intradimensional Bias. Moreton (2012) showed that, when learning phonotactics, subjects trained on patterns that involved the same feature across phonemes (i.e., intradimensional ones) had higher accuracy at the end of learning than subjects learning patterns that used two different features across the same number of sounds. Moreton (2012) suggested that these findings could explain why patterns like assimilation are more typologically common than patterns that aren't intradimensional.

The first simulations I'll describe deal with Moreton's (2012) Height-Height (HH) and Height-Voicing (HV) languages, since those are the two he focused on in his own simulations. In the former pattern, the two vowels in a word must match in their value for the feature [High],

while in the latter pattern, high vowels are always followed by voiced consonants and low vowels are always followed by voiceless ones.

To represent the words in these languages, I used the phonemes and corresponding feature values shown in Table 9. The training data consisted of every possible combination of these sounds with a length of two (since there are only two relevant sounds in each pattern). In the HH language, every word with two vowels that matched in their value for [High] (N=8) mapped to a probability of 0.125 (1/8), with the rest of the possible bigrams mapping to probabilities of 0. For the HV language, every word that contained either a high vowel followed by a voiced consonant, or a low vowel followed by a voiceless consonant (N=8) was given a probability of 0.125 (1/8), while all other words mapped to 0.

	[voice]	[dorsal]	[high]	[back]
t	-	-	-	-
d	+	-	-	-
k	-	+	-	-
g	+	+	-	-
i	-	-	+	-
u	-	-	+	+
a	-	-	-	-
o	-	-	-	+

Table 9. Features and phonemes used in simulating the results from Moreton (2012).

First, the standard version of GMECCS was trained on each language for 200 epochs, with a learning rate of 0.01. The average learning curves over 15 runs for each of the languages are shown in Figure 7.

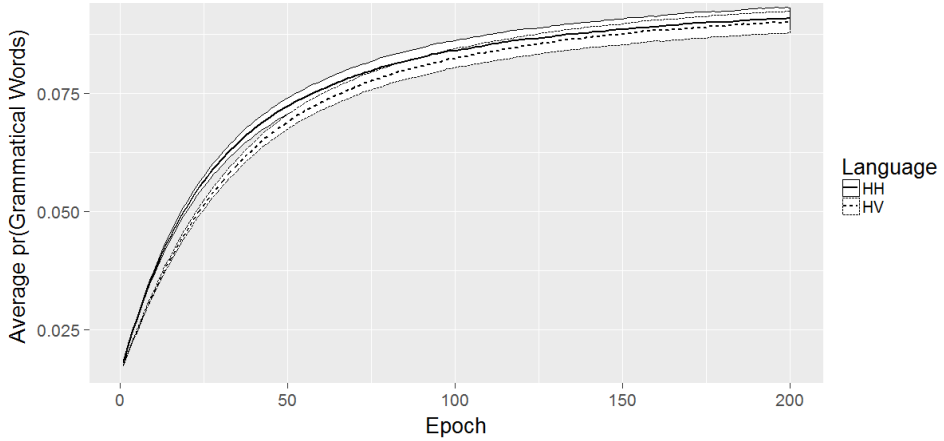

Figure 7. Learning curve for the model, when trained on the Moreton (2012) HH and HV patterns with a version of GMECCS that has no variables and does not use PFA. White space behind each line shows the standard error of the mean at that epoch.

GMECCS has a small preference for the HH language, but rather than being the result of a consistent preference for intradimensional patterns, this is simply due to the fact that the HH pattern has no legal words that contain consonants. GMECCS is able to more quickly move probability mass onto the correct words, because simple constraints like "*[+Dorsal]" and "*[+Voice]" quickly receive higher weights early on in the learning process. To demonstrate this, I ran simulations for another one of Moreton's (2012) patterns: Height-Backness (HB). This pattern is similar to HV in that it involves a dependency across two different features in two different segments. However, in HB only vowel features are relevant to the pattern, so it isn't necessary to include any consonants and the confound present in the simulations described above can be avoided. Results comparing the model without PFA that was trained on HB and HH, averaged over 15 repetitions, are shown in Figure 8.

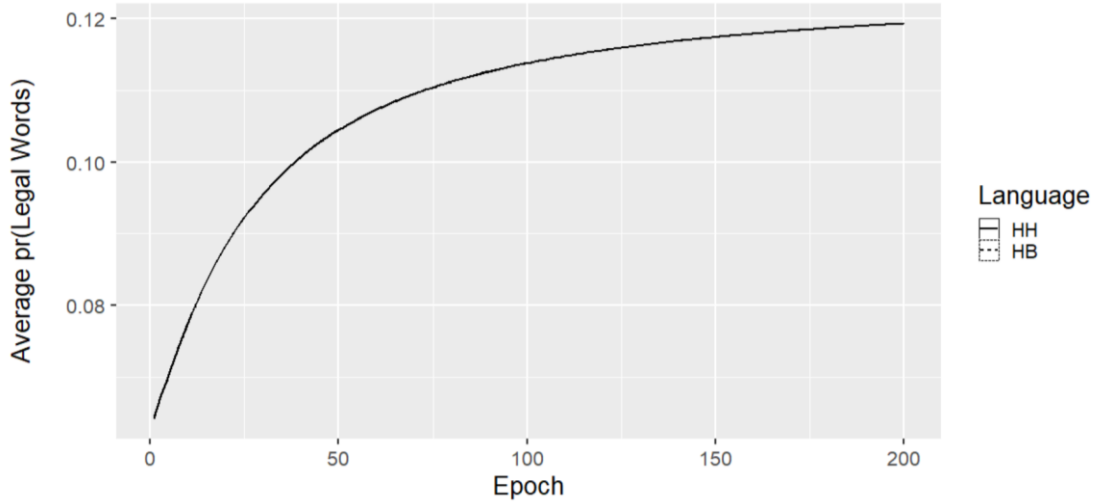

Figure 8. Learning curve for the model, when trained on the Moreton (2012) HH and HB patterns with a version of GMECCS that has no variables and does not use PFA. Standard error of the mean at each epoch is too small to be shown.

This figure confirms that GMECCS's HH>HV preference was due to the former pattern's lack of consonants and not an Intradimensional Bias. This represents another effect that Moreton (2012) found—a bias towards patterns on a "single tier" (see, for example, his Table 8). If GMECCS was able to capture an Intradimensional Bias, it would also have a HH>HB preference, but these two languages are treated identically by the model. This failure is a result of the fact that GMECCS has no way of representing inter-segmental similarity in its constraints and means that, unlike humans, GMECCS is blind to the differences between the HH and HB

patterns. Moreton (2012) suggests variables as a way to capture the necessary bias (as described in §2.1), but PFA can do it without any algebraic machinery.

To demonstrate this, I ran simulations with the exact same patterns and hyperparameters as above, but with a model using PFA. The results for this, averaged over 15 separate runs are shown in Figure 9.

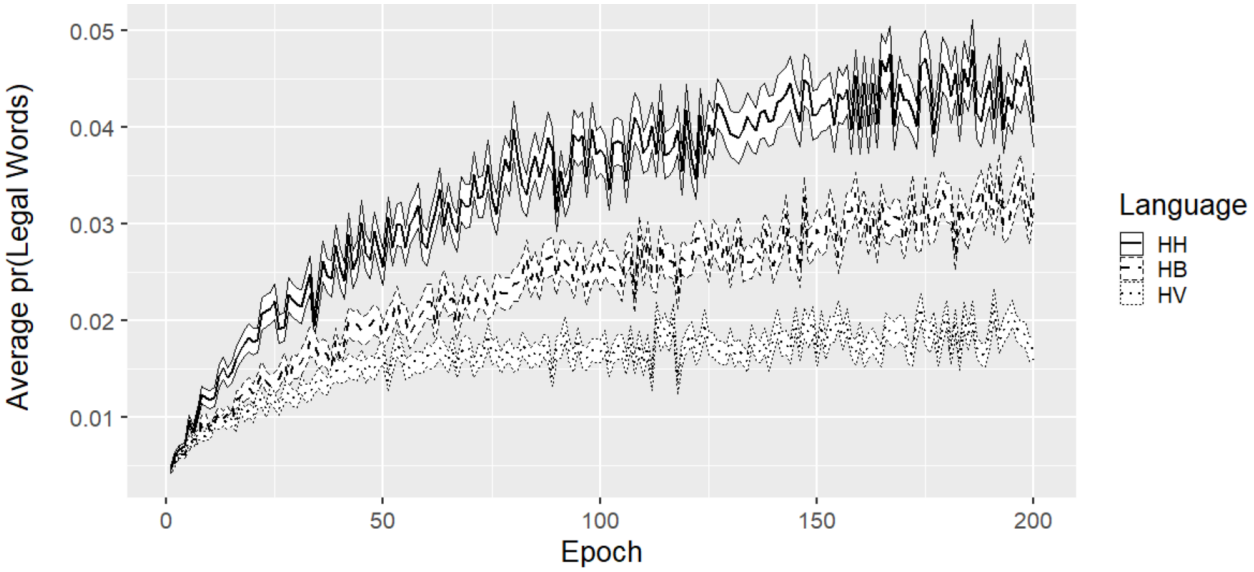

Figure 9. Learning curve for the model, when trained on the Moreton (2012) Experiment 1 data with a version of GMECCS that uses PFA. White space behind each line shows the standard error of the mean at that epoch.

These results demonstrate that when PFA is used, probability is moved onto the correct words more quickly in the HH language than in the HV *or* HB languages.[14] This is a result of the fact that each weight update is likely to be using a subset of the possible features available to the model. So, the more features that are necessary for any individual pattern (within or between segments), the more likely that the pattern will be obscured due to PFA. Since HH only involves one feature and HV/HB both involve two, the latter patterns are more difficult for a model with PFA to acquire. This means that PFA captures Intradimensional Bias without explicitly representing inter-segmental similarity through algebraic variables, and like Identity Bias, this success is independent of the exact feature theory or data set being used.

4.4. Similarity-based Generalization

This section focuses on a phenomenon that variables cannot account for: *Similarity-based Generalization*, which is the term I'll use to describe the process of a phonological pattern being

[14] Like the vanilla version of GMECCS, the model also has the "single tier" effect that Moreton (2012) found in his participants: patterns that involve only vowels (like HH and HB) are easier to learn than those that don't (like HV).

generalized to novel words in a way that is predicted better by simple similarity metrics than by theories of phonological generalization that rely on feature-based classes. For example, Halle (1978) suggested that only novel sounds that are described by the simplest featural description of a pattern (often called a *natural class*) should be generalized to. For example, if a person were learning a language where all words began with [+Voice] sounds, they would be predicted to only generalize this pattern to words beginning with novel [+Voice] sounds, and not to any words that began with [−Voice] ones.

However, humans seem to sometimes generalize phonological processes and restrictions outside of the natural classes that they originally applied to. This kind of generalization has been argued to affect diachronic changes in natural language (Mielke 2008 sec. 5.2.2) and was also observed by Cristia et al. (2013) in an artificial language learning experiment. After being exposed to a voicing restriction in the onset of words, Cristia et al.'s (2013) participants were asked to rate the acceptability of novel stimuli that belonged to one of three categories: WITHIN words, which began with sounds that belonged to the same natural class as those that were present in training, NEAR words, whose initial sound was relatively similar[15] to those in training (but did not belong to the same natural class), and FAR words, which began with sounds that were dissimilar to those in the experiment's training phase. For example, when trained on words that began with all of the voiced sounds in English except [b], participants were tested on their generalization to words that began with [b] (WITHIN), [k] (NEAR), and [p] (FAR). The sound [k] was considered NEAR because [g]-initial words were present in training and only differed from [k] in their value for [Voice]. However, [p] was FAR because no labial stops or voiceless sounds were present in the experiment's training phase.

Standard theories of phonological generalization would predict higher acceptability ratings for WITHIN words than for words in the other two categories. However, participants' ratings revealed that they not only generalized the restriction to the WITHIN words, but also to the NEAR ones. FAR, on the other hand, had significantly lower ratings than the other two groups. These experiment results, along with participant ratings for items that were present in training (labeled "Exposure") are shown in Figure 10, adapted from Cristia et al.'s (2013) *Figure 3*.

[15] NEAR words were similar to those in training along three dimensions: raw acoustic measurements, raw articulatory measurements, and the average number of phonological features that the sounds differed in.

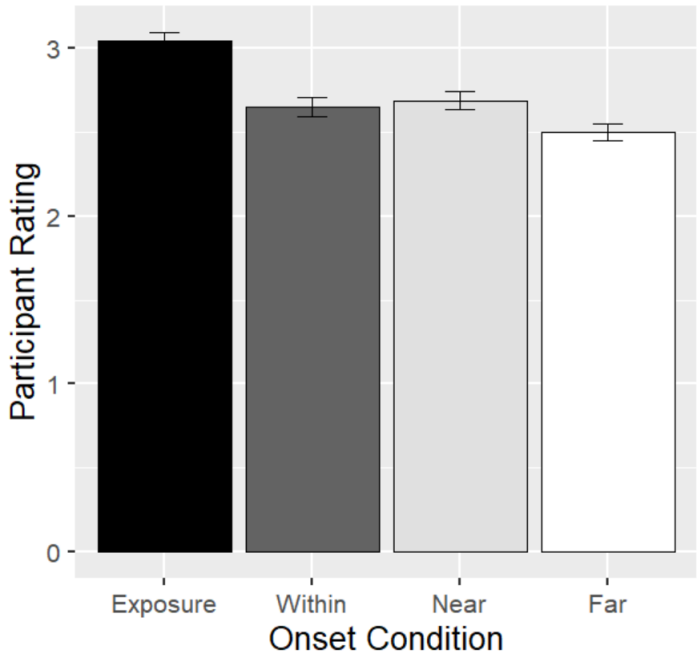

Figure 10. Human results adapted from Cristia et al.'s (2013) Figure 3. Error bars show standard error of the mean and are calculated using data from the paper's supplementary materials.

To see whether the standard version of GMECCS is able to model this phenomenon, I ran a simulation of the Cristia et al. (2013) experiment using the phonemes and features given in Table 10. Ten separate runs were performed for each of Cristia et al.'s (2013) 12 experiment conditions.[16] In each condition, words were simplified to just their initial consonant (i.e. the only sound that was relevant to the voicing restriction).

	[labial]	[dorsal]	[voice]	[palatal]	[continuant]
p	+	-	-	-	-
b	+	-	+	-	-
f	+	-	-	-	+
v	+	-	+	-	+
t	-	-	-	-	-
d	-	-	+	-	-
s	-	-	-	-	+
z	-	-	+	-	+
ʃ	-	-	-	+	+
ʒ	-	-	+	+	+
k	-	+	-	-	-
g	-	+	+	-	-

Table 10. Features and phonemes used in simulating the results from Cristia et al. (2013).

[16] Each condition in Cristia et al.'s (2013) experiment represented a different possible set of WITHHELD, NEAR, and FAR segments. Six conditions restricted word-initial consonants to being voiced sounds and six restricted them to being voiceless.

Training data consisted of every possible consonant mapping to a probability of either .2 (i.e., 1/5) or 0. Attested words (i.e. those that mapped to .2) all consisted of either a [+Voice] or [−Voice] sound (depending on the condition). The standard version of GMECCS was trained on this data for 200 epochs, with a learning rate of 0.01. The model's probability estimates at the end of learning (which were representative of the entire acquisition process) for the different groups of testing data, as well as data from training (labeled "Exposure") are shown in Figure 11.

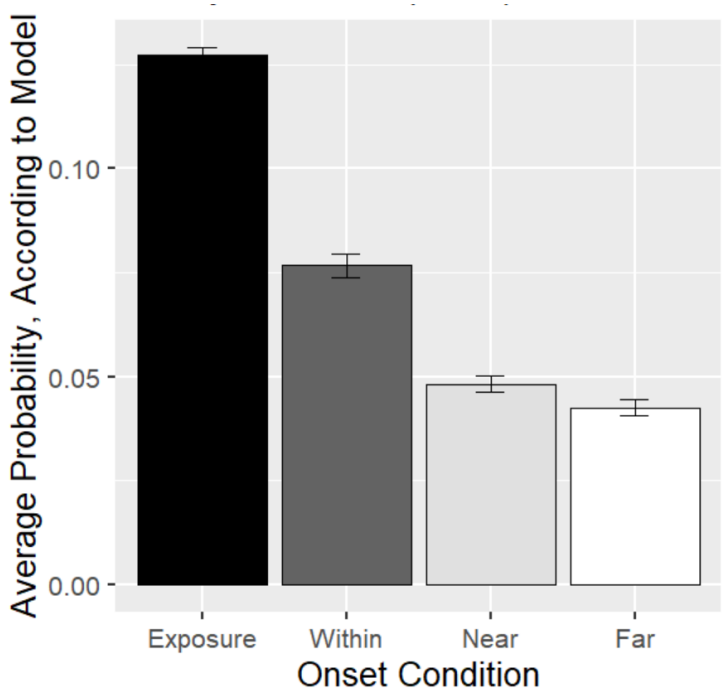

Figure 11. Probabilities for the Cristia et al. (2013) testing data, after 200 epochs, with a version of GMECCS that has no variables and does not use PFA. Error bars show standard error of the mean.

Without any added mechanisms, GMECCS fails to capture Similarity-based Generalization. NEAR and FAR test items are both given relatively low probability, with the model estimating a much higher probability for the WITHIN category. This is because the model's constraints can easily represent the feature-based grouping present in the training data. For example, when attested words are all voiced, the constraint "*[−Voice]" will receive a relatively high weight. This means that even unattested [+Voice] sounds (i.e. sounds in the WITHIN condition) will be given a relatively high probability. However, there is no single constraint that groups segments in the NEAR condition with all the attested sounds, meaning that these data receive relatively low probability estimates.

Variables cannot be used to address this prediction of the model because they must be restricted to occurring on the same feature across different segments. While this restriction was not present in Halle's (1962) original proposal for variables, it has a number of theoretical (McCawley 1971, Schuh 1978, Odden 2013) and empirical motivations (Moreton 2012, Prickett 2020). For example, Moreton (2012) points out that the Intradimensional Bias that his model captures using variables would disappear if the variables could be used to simplify the representation of non-assimilatory patterns. The standard way of ensuring this doesn't happen, and the method that Moreton (2012) follows, is the above-mentioned restriction. This also means that there is no way that variables could affect the learning or generalization of a single-segment pattern like the one Cristia et al. (2013) used, since variables that are restricted in this way can only be used to represent similarity across segments.

However, PFA *can* capture this behavior. I ran the same simulations described above with the version of GMECCS that uses PFA. The results for this after 200 epochs (a somewhat representative point in learning[17]) are shown in Figure 12.

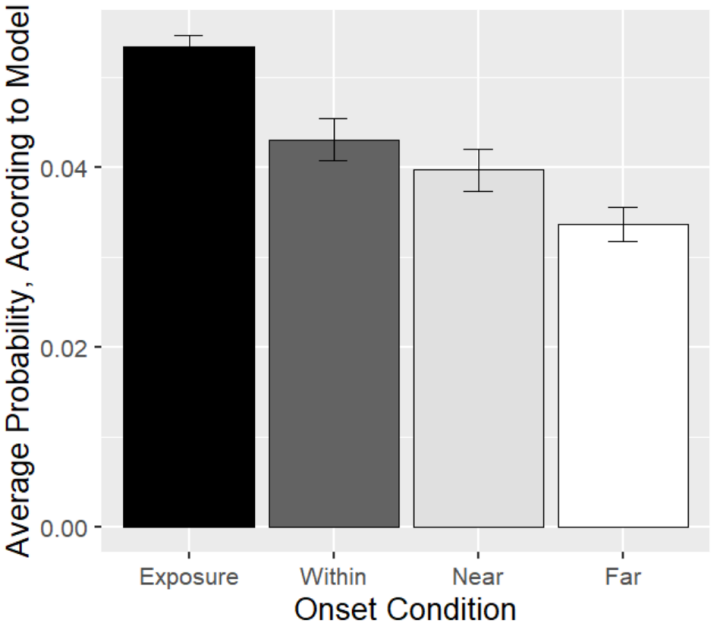

Figure 12. Probabilities for the Cristia et al. (2013) testing data, after 200 epochs, with a version of GMECCS that uses PFA. Error bars show standard error of the mean.

These results show that PFA is able to capture Similarity-based Generalization, since there is only a marginal difference between the probability given to the WITHIN and NEAR items, but

[17] While the same qualitative results exist throughout learning for the PFA model, the standard error of the mean for the WITHIN and NEAR conditions only starts overlapping after about 125 epochs.

both groups have a consistently higher probability than FAR.[18] This is because the more featurally similar two sounds are, the more likely they are to be ambiguous with one another on any given weight update. Since NEAR words are, by definition, more featurally similar to the training items than FAR words, they are more likely to become ambiguous with words that map to high probabilities in the training data.

For example, if the model is trained to map [g] to a probability of .2, but [k] and [p] to a probability of 0, on unambiguous trials, it will raise the weights of constraints that [k] and [p] violate, such as "*[−Voice, +Dorsal]" and "*[−Voice, +Labial]", respectively. However, if [Voice] is not attended to, the model could be presented with the datum [?Voice, +Dorsal] mapping to a probability of .2 (i.e. [g] with ambiguous voicing). This would cause the model to lower the weights of both "*[+Voice, +Dorsal]" *and* "*[−Voice, +Dorsal]", thus increasing the probability it assigns to both [g] and NEAR sounds like [k]. A scenario like this is less likely to happen with [p], since two features would need to be ignored for it to be ambiguous with words that map to nonzero probabilities.

This results in the model predicting the kind of Similarity-based Generalization that Cristia at al. (2013) observed in their participants' behavior and like the phenomena discussed in §4.2 and §4.3, is independent of whatever feature theory or data set the model happens to be given.

5. Discussion

5.1.Future work

There are a number of ways in which PFA can be extended and further examined. Applying PFA to more realistic language data, like Berent et al.'s (2012) work with Identity-based Generalization in Hebrew or Gallagher's (2013) computational work with Chol, could help discover whether more complex data in training changes PFA's effects. While the natural language examples are formally similar to the artificial languages tested here, the larger number of features and segments could cause ambiguity to push the learner in unexpected directions.

While real language data could be difficult to handle with the large number of constraints that GMECCS requires, equipping the Hayes and Wilson (2008) model with PFA could be a more efficient way to test this question (alternatively, other methods of constraint induction

[18] The model also predicts a considerable amount of variation across experiment conditions. Since Cristia et al. (2013) only had two participants in each condition, it's impossible to know whether this aspect of the model's predictions is accurate without a replication of the experiment that involves more participants.

could be used such as those proposed in Pizzo 2013, Moreton 2019). Another more efficient alternative to GMECCS could be recurrent neural networks. These have successfully been applied to learning phonotactic patterns with real language data (see, e.g., Mirea & Bicknell 2019, Mayer & Nelson 2020) and Prickett et al. (2018) showed that dropout, a mechanism similar to PFA, could assist a neural network in performing Identity-based Generalization.

A number of decisions had to be made about how to implement PFA, and future work should see which of these have an effect on the mechanism's usefulness. For example, in the simulations presented here, features were either completely attended to or completely ignored. But more gradient weights could be randomly assigned to features instead,[19] which would make PFA more closely resemble past approaches to featural attention (Nosofsky 1986). Additionally, the features that were used were all binary, but PFA could be applied to more complex feature geometries to see what it might predict in that scenario.

The results in this paper were all modeling adult language learning. However, another avenue for future work would be comparing the predictions that PFA and variables make regarding child language acquisition. If children make the kind of mistakes predicted by either variables or PFA, it could be strong evidence for the respective theory. Modeling experiments like Gervain and Werker (2013), which explore how infants' ability to learn identity-based patterns changes over time, could be a good place for this line of research to start.

Broadly speaking, the results in §4.1 and §4.4 show that PFA increases the likelihood that a model will generalize in a human-like way to novel data—a job typically carried out in machine learning by adding mathematical priors to an objective function. Future work should explore the relationship between PFA and other forms of regularization, since this could be useful in understanding its mathematical implications (see Wager, Wang, & Liang 2013 for more on how dropout is equivalent to a special case of L2 priors).

5.2.Conclusions

Previous literature on whether variables are necessary in models of phonotactic learning used a number of human behaviors as evidence for their existence (Berent et al. 2012, Moreton 2012, Gallagher 2013, Berent 2013). Here I have shown that three of these phenomena can be

[19] Thanks to an anonymous reviewer for suggesting this possibility.

explained using Probabilistic Feature Attention: Identity-based Generalization, Identity Bias, and Intradimensional Bias.

This simple mechanism, which relies primarily on the assumption that learners may not always attend to every phonological feature in a surface form, offers a unified account of not only the variable-related phenomena, but also Similarity-based Generalization, something that variables cannot explain. It does this by introducing ambiguity across forms throughout learning, which causes forms that are featurally similar to one another to be treated similarly by the model. While PFA predicts that only certain sets of data and feature-based representations will lead to Identity-Based generalization, I found that it predicts the other three phenomena in a way that is independent of these assumptions.

These results suggest that Probabilistic Feature Attention, paired with a variable-free model of phonotactics, could be a viable alternative to theories of phonological learning that make use of explicit variables in their representations.

References

Alhama, Raquel G., & Zuidema, Willem. (2018). Pre-Wiring and Pre-Training: What does a neural network need to learn truly general identity rules? *Journal of Artificial Intelligence Research 61* 927–946.

Bailey, Todd M., & Hahn, Ulrike. (2005). Phoneme similarity and confusability. *Journal of Memory and Language 52*(3) 339–362. https://doi.org/10.1016/j.jml.2004.12.003

Berent, Iris. (2013). The phonological mind. *Trends in Cognitive Sciences 17*(7) 319–327.

Berent, Iris, Marcus, Gary, Shimron, Joseph, & Gafos, Adamantios I. (2002). The scope of linguistic generalizations: Evidence from Hebrew word formation. *Cognition 83*(2) 113–139.

Berent, Iris, & Shimron, Joseph. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition 64*(1) 39–72.

Berent, Iris, Wilson, Colin, Marcus, Gary, & Bemis, Douglas K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry 43*(1) 97–119.

Chomsky, Noam, & Halle, Morris. (1968). *The sound pattern of English.* Harper & Row.

Corina, David Paul. (1991). *Towards an understanding of the syllable: Evidence from linguistic, psychological, and connectionist* [PhD Thesis]. University of California, San Diego.

Cristia, Alejandrina, Mielke, Jeff, Daland, Robert, & Peperkamp, Sharon. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology 4*(2) 259–285.

Doucette, Amanda. (2017). Inherent Biases of Recurrent Neural Networks for Phonological Assimilation and Dissimilation. *ArXiv Preprint ArXiv:1702.07324*.

Gallagher, Gillian. (2013). Learning the identity effect as an artificial language: Bias and generalisation. *Phonology 30*(2) 253–295.

Gasser, Michael. (1993). *Learning words in time: Towards a modular connectionist account of the acquisition of receptive morphology*. Indiana University, Department of Computer Science.

Gasser, Michael, & Lee, Chan-Do. (1992). Networks that learn about phonological feature persistence. In *Connectionist Natural Language Processing* (pp. 349–362). Springer.

Gathercole, Susan E., & Adams, Anne-Marie. (1993). Phonological working memory in very young children. *Developmental Psychology 29*(4) 770–778. https://doi.org/10.1037/0012-1649.29.4.770

Gervain, Judit, & Werker, Janet F. (2013). Learning non-adjacent regularities at age 0; 7. *Journal of Child Language 40*(4) 860–872.

Gluck, Mark A., & Bower, Gordon H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language 27*(2) 166–195.

Goldsmith, John. (1976). *Autosegmental phonology*. MIT Press London.

Greenberg, Joseph H. (1950). The patterning of root morphemes in Semitic. *Word 6*(2) 162–181.

Halle, Morris. (1962). A descriptive convention for treating assimilation and dissimilation. *Quarterly Progress Report 66* 295–296.

Halle, Morris. (1978). *Knowledge unlearned and untaught: What speakers know about the sounds of their language*.

Hare, Mary. (1990). The role of trigger-target similarity in the vowel harmony process. *Annual Meeting of the Berkeley Linguistics Society 16* 140–152.

Hayes, Bruce. (2011). *Introductory phonology* (Vol. 32). John Wiley & Sons.

Hayes, Bruce, & Wilson, Colin. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*(3) 379–440.

Jarosz, Gaja. (2013). Learning with hidden structure in optimality theory and harmonic grammar: Beyond robust interpretive parsing. *Phonology 30*(1) 27–71.

Linzen, Tal, & Gallagher, Gillian. (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology 8*(1).

Marcus, Gary. (1999). Do infants learn grammar with algebra or statistics? Response. *Science 284*(5413) 436–437.

Marcus, Gary. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, Gary, Vijayan, Sugumaran, Rao, S. Bandi, & Vishton, Peter M. (1999). Rule learning by seven-month-old infants. *Science 283*(5398) 77–80.

Mayer, Connor, & Nelson, Max. (2020). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics 3*(1) 149–159.

McCarthy, John J. (2010). Autosegmental spreading in Optimality Theory. In John A. Goldsmith, Elizabeth Hume, & W. Leo Wetzels (Eds.) *Tones and Features* (pp. 195–222). Walter de Gruyter.

McCawley, James D. (1971). On the role of notation in generative phonology. In Maurice Gross, Morris Halle, & Marcel-Paul Schützenberger (Eds.) *The formal analysis of natural languages* (pp. 51–62). Mouton, The Hague.

Mielke, Jeff. (2008). *The emergence of distinctive features*. Oxford University Press.

Mirea, Nicole, & Bicknell, Klinton. (2019). Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 1595–1605.

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

Moreton, Elliott. (2019). Constraint breeding during on-line incremental learning. *Proceedings of the Society for Computation in Linguistics 2*(1) 69–80.

Moreton, Elliott, Pater, Joe, & Pertsova, Katya. (2017). Phonological Concept Learning. *Cognitive Science 41*(1) 4–69.

Nosofsky, Robert M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General 115*(1) 39.

Odden, Dave. (2013). Formal phonology. *Nordlyd 40*(1) 249–273.

Pater, Joe, & Moreton, Elliott. (2014). Structurally biased phonology: Complexity in learning and typology. *The EFL Journal 3*(2).

Pizzo, Presley. (2013). *Learning Phonological Alternations with Online Constraint Induction*. The 10th Old World Conference in Phonology Istanbul. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.705.4286&rep=rep1&type=pdf

Prickett, Brandon. (2020). Variables Must be Limited to a Single Feature. *Proceedings of the Annual Meetings on Phonology 8.*

Prickett, Brandon, Traylor, Aaron, & Pater, Joe. (2018). Seq2Seq Models with Dropout can Learn Generalizable Reduplication. *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology* 93–100.

Rabagliati, Hugh, Ferguson, Brock, & Lew-Williams, Casey. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science 22*(1) e12704.

Rose, Sharon, & Walker, Rachel. (2004). A Typology of Consonant Agreement as Correspondence. *Language 80*(3) 475–531.

Schuh, Russell G. (1978). Tone rules. In *Tone* (pp. 221–256). Elsevier.

Smolensky, Paul, & Legendre, Géraldine. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, & Salakhutdinov, Ruslan. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*(1) 1929–1958.

Štekauer, Pavol, Valera, Salvador, & Körtvélyessy, Lívia. (2012). *Word-formation in the world's languages: A typological survey*. Cambridge University Press.

Tang, Kevin, & Baer-Henney, Dinah. (2019). *Disentangling L1 and L2 effects in artificial language learning*. Manchester Phonology Meeting Manchester, UK. http://www.lel.ed.ac.uk/mfm/27mfm-abbk.pdf

Tesar, Bruce. (2004). Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry 35*(2) 219–253.

Wager, Stefan, Wang, Sida, & Liang, Percy S. (2013). Dropout training as adaptive regularization. *Advances in Neural Information Processing Systems* 351–359.