Assessing the learnability of process interactions using grammatical spaces

Christopher Yang
Linguistics and Philosophy
MIT
cminwoo@mit.edu

Adam Albright
Linguistics and Philosophy
MIT
albright@mit.edu

Naomi H. Feldman
Linguistics and UMIACS
University of Maryland
nhf@umd.edu

Abstract

A challenge in learning phonological grammars is learning how phonological processes interact. It has been argued that some process interactions are easier to learn than others. One basis for this argument is asymmetries observed in experimental settings: artificial languages generated from certain process interactions are more likely to be successfully reproduced by participants than others. In this paper, we argue that asymmetries in production do not necessarily provide direct support that some phonological interactions are easier to learn. Rather, we show that these asymmetries can instead emerge due to differences in the number of consistent or nearly-consistent grammars each pattern has. We present a noisy channel model of morpho-phonological learning and apply it to a recent behavioral study examining the learnability of phonological process interactions. We find that, due to the relative difference in the number of grammars that can exactly match or nearly match the observed data, the model achieves the same qualitative results as those observed in experimental settings.

Keywords: Linguistics; Phonology; Opacity; Noisy Channel; Bayesian

Introduction

A basic empirical observation is that languages change. An example of this is given by Kiparsky (1968, 1971), who argued that patterns generated from certain phonological process interactions are more likely to change over time than others. This asymmetry served as motivation to distinguish process interactions based on learnability: learners are more likely to successfully learn certain process interactions (i.e. transparent interactions) than others (i.e. opaque interactions). Recent work has sought to experimentally and computationally investigate these learnability claims by evaluating participants' and models' performance in artificial grammar learning paradigms (Ettlinger, 2008; Kim, 2012; Brooks, Pajak, & Baković, 2013; Prickett, 2019). These studies found evidence that data generated by some process interactions were more difficult to reproduce than others.

The standard interpretation of these results is that if a pattern is easier to reproduce, then the interaction used to generate that pattern is also easier to learn. However, we argue that this conclusion is premature; there are in fact many ways of generating the same set of data beyond learning the grammar that generated it. These include memorizing parts of the data to hypothesizing entirely different phonological processes than those of the original grammar. In this paper, we explicitly show how considering the *space* of grammars is capable of capturing the reported learnability asymmetries. To this end, we implement a noisy-channel learner that jointly infers underlying forms and phonological processes from form-meaning output pairs. The model generates predictions given a space of (nearly) consistent grammars, which we compare with empirical results from a recent behavioral study by Prickett (2019). We find that this model, due to the difference in the number of grammars that can generate or nearly generate different languages, achieves the same qualitative results as those observed in experimental settings.

We begin first with some background on process interactions. We then present a recent behavioral study investigating the learnability of process interactions, as well as the artificial languages we will be using. Next, we provide an overview of the model and compare the model predictions with the empirical results. We conclude with implications and next steps.

Background

One of the goals in generative phonology is characterizing how infants acquire a phonological grammar from a set of surface representations (SR; e.g. words). A phonological grammar is comprised of two parts: a *lexicon* containing the representations (also known as the underlying representation, or UR) of atomic meanings and the *mapping* from URs to observed surface forms. For example, many languages delete a vowel when it precedes another vowel[1]:

Etsako (Casali, 1997) (1)
[dɛ] 'buy' + [akpa] 'cup'
[d # akpa] 'buy (a) cup'

When [ɛ] does and does not appear is predictable based on its context: [ɛ] does not surface when before another vowel. We can generate each output given a UR and a context-sensitive mapping that transforms the UR into its observed form: /dɛ/ → [dɛ] vs. /dɛ # akpa/ → [d # akpa].

In addition, in many languages, consonants undergo palatalization in certain contexts (typically, before non-back (high) vowels like [i] or glides like [j]). Again, the distribution of [t] and [tʃ] given the data is predictable based on

[1] In reality, this characterization is not restrictive enough. The vowel that gets deleted in this language depends on a variety of factors, including which vowels are adjacent to each other. We will set aside these complications for the time being.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Table 1: Sample toy languages for the bleeding and feeding interactions. Shaded cells correspond to filler words. Filler words were not used to evaluate participants in either the experiment or the model.

Bleeding	*Stem-a*	*Stem-i*	*Stem-i-a*
UR	/imat-a/	/imat-i/	/imat-i-a/
Deletion	–	–	[imat-_-a]
Palatalization	–	[imatʃ-i]	–
SR	[imat-a]	[imatʃ-i]	[imat-_-a]
	Stem-a	*Stem-i*	*Stem-i-a*
UR	/imak-a/	/imak-i/	/imak-i-a/
Deletion	–	–	[imak-_-a]
Palatalization	–	–	–
SR	[imak-a]	[imak-i]	[imak-_-a]

Feeding	*Stem-a*	*Stem-i*	*Stem-i-a*
UR	/imat-a/	/imat-i/	/imat-a-i/
Deletion	–	–	[imat-_-i]
Palatalization	–	[imatʃ-i]	[imatʃ-_-i]
SR	[imat-a]	[imatʃ-i]	[imatʃ-_-i]
	Stem-a	*Stem-i*	*Stem-i-a*
UR	/imak-a/	/imak-i/	/imak-a-i/
Deletion	–	–	[imak-_-i]
Palatalization	–	–	–
SR	[imak-a]	[imak-i]	[imak-_-i]

context: [tʃ] appears before [i], while [t] appears elsewhere.

Japanese (2)
[mot-e] 'hold-IMP'
[mot-anai] 'hold-NEG'
[motʃ-imas-u] 'hold-PRES-POL'

We characterize the mapping from inputs to outputs via a rule-based approach (SPE; Chomsky & Halle, 1968). Under this approach, mappings are computed through the sequential application of context-sensitive rules; if a sound is found in a particular context, that sound is changed. For example, we can formally characterize the palatalization and deletion processes using the following notation:

Palatalization: t → tʃ / _i (3)
Deletion: V → Ø / _V

The first process states that if [t] is found before [i], the sound changes into [tʃ]. The second process states that if a vowel is found before another vowel, the first vowel deletes.

Process interactions. Individual processes can potentially interact with one another. There are two dimensions along which these interactions are characterized under a rule-based system: whether a process creates or eliminates the string to which another process applies, and whether the process precedes or follows the other process. For example, consider the UR /imat-a-i/ and the processes in (3). If deletion precedes palatalization, the application of deletion creates the relevant context in which palatalization can apply: /imat-a-i/ $\rightarrow_{DEL}$ [imat-_-i] $\rightarrow_{PAL}$ [imatʃ-_-i]. This is known as a *feeding* interaction. If the opposite ordering were to occur, the deletion process would have applied too late to trigger palatalization, i.e. [imat-_-i]. This is known as a *counter-feeding* interaction.

Consider now a slightly different UR /imat-i-a/. The only difference between this input and the preceding example is that the order of the suffixes is swapped. If deletion precedes palatalization, deletion now *eliminates* the environment in which palatalization could have applied: /imat-i-a/ $\rightarrow_{DEL}$ [imat-_-a] $\rightarrow_{PAL}$ [imat-_-a]. This is known as a *bleeding* interaction. Reversing the order of the processes would allow the application of palatalization to occur before deletion can eliminate the context, i.e. [imatʃ-_-a]. This is known as a *counter-bleeding* interaction.

Note that rule-based systems are not the only way of formalizing this mapping. More recent approaches characterize the mapping though the evaluation and elimination of possible outputs via ranked or weighted constraints (Prince & Smolensky, 2004; Goldwater & Johnson, 2003). We opt for the rule-based approach for two reasons. First, most constraint-based formalisms lack the necessary mechanisms to represent the crucial process interactions. Second, the rule-based approach does not have an intrinsic bias in favor of any particular phonological process. Minimizing these prior biases in the model ensures that any asymmetry observed in performance can be attributed to the grammatical space.

Toy languages. Learners acquiring the phonology of a language need to learn a set of URs and rules. The question of which types of process interactions are easiest to learn, and how this relates to cross-linguistic patterns, has attracted considerable attention. The focus of our simulations is a recent study by Prickett (2019). Prickett investigated the learnability of different process interactions in an experimental setting. He trained participants in one of the four toy languages, associated with each of the process interactions discussed above.

A toy language is composed of morpho-phonological paradigms. Each paradigm consists of the stem (e.g. [imat]) in isolation, as well as three conjugated forms: the stem with an [-i] suffix, the stem with an [-a] suffix, and the stem with both the [-i] and [-a] suffixes. The order in which the [-i] and [-a] suffixes combine differs depending on the language. Each language consists of two paradigm types: [t]-final stems (e.g. /imat/) and [k]-final stems (e.g. /imak/).

Each slot in the paradigm is generated by sequentially applying the deletion and palatalization processes to each stem-suffix combination. For the bleeding and feeding languages, each UR undergoes deletion followed by palatalization. For the counter-bleeding and counter-feeding languages, each UR undergoes palatalization followed by deletion. The four languages are identical in all respects except for when both suffixes attach to the stem; for example, the counter-bleeding interaction generates [imatʃ-_-a], while the counter-feeding interaction generates [imat-_-i]. Sample paradigms for the

Table 2: Trial types and schematized choices for each toy language. Expected responses given on the left in bold. Cells in the table correspond to the same cells in Table 1.

Bleeding	*Faithful*	*Palatalizing*	*Interacting*
SR	[imat-a]	[imatʃ-i]	[imat-_-a]
Options	**[t-a]** *[tʃ-a]	**[tʃ-i]** *[t-i]	**[t-_-a]** *[tʃ-_-a]
	–	–	*Deleting*
SR	–	–	[imak-_-a]
Options	–	–	**[k-_-a]** *[k-i-a]

Feeding	*Faithful*	*Palatalizing*	*Interacting*
SR	[imat-a]	[imatʃ-i]	[imatʃ-_-i]
Options	**[t-a]** *[tʃ-a]	**[tʃ-i]** *[t-i]	**[tʃ-_-i]** *[t-_-i]
	–	–	*Deleting*
SR	–	–	[imak-_-i]
Options	–	–	**[k-_-i]** *[k-a-i]

Counter-bleeding	*Faithful*	*Palatalizing*	*Interacting*
SR	[imat-a]	[imatʃ-i]	[imatʃ-_-a]
Options	**[t-a]** *[tʃ-a]	**[tʃ-i]** *[t-i]	**[tʃ-_-a]** *[t-_-a]
	–	–	*Deleting*
SR	–	–	[imak-_-a]
Options	–	–	**[k-_-a]** *[k-i-a]

Counter-feeding	*Faithful*	*Palatalizing*	*Interacting*
SR	[imat-a]	[imatʃ-i]	[imat-_-i]
Options	**[t-a]** *[tʃ-a]	**[tʃ-i]** *[t-i]	**[t-_-i]** *[tʃ-_-i]
	–	–	*Deleting*
SR	–	–	[imak-_-i]
Options	–	–	**[k-_-i]** *[k-a-i]

bleeding and feeding languages are given in Table 1.

We refer to certain forms of each toy language using terminology that reflects which process was being tested. Faithful forms correspond to the [t]-final stem combined with the [-a] suffix /imat-a/. Palatalizing forms correspond to the [t]-final stem combined with the [-i] suffix /imat-i/. Deleting forms correspond to the [k]-final stem combined with both the [-i] and [-a] suffixes /imak-a-i/ and /imak-i-a/. Lastly, the interacting forms correspond to the [t]-final stem combined with both the [-i] and [-a] suffixes /imat-a-i/ and /imat-i-a/.

The training phase consisted of giving the stem in isolation to the participant, then having them produce some suffixed form by selecting one of two options. For example, in the feeding language, the participant first hears the word [imat] paired with an image representing its meaning. The participant is then given an image representing the morphologically complex form containing the [-a] suffix and asked to choose between the options [imat-a] and [imatʃ-a]. Feedback is then given as to whether the response is correct or not. The testing phase was identical to the training phase except that participants were presented novel words and feedback was not given. Prickett tracked the proportion of responses by each speaker that was or was not predicted under the intended grammar. Schematizations of the choices and expected responses for each trial type are given in Table 2. The results of the experiment are shown on the top of Figure 1.

Prickett found two statistically significant results. In the palatalizing trials, Prickett found that participants trained on the feeding and counter-bleeding languages had significantly higher accuracy than the bleeding and counter-feeding languages. In the interacting trials, Prickett found that participants performed better when generating the bleeding and feeding versus the counter-bleeding and counter-feeding forms. It is tempting to interpret the results of this experiment as providing direct evidence that the process interactions that yielded high accuracy are easiest to learn. However, we argue in this paper that such a conclusion would be premature.

Grammatical spaces. While the outputs were *generated* via the application of particular phonological processes for a defined UR, this exact grammar is not the only manner in which the data could have been produced. The language could have also be generated through the application of different phonological processes, through memorization, or a combination of the two. For example, the interacting form for the counter-feeding interaction [imat-_-i] could have been generated by having its UR be /imat-_-i/ with neither palatalization nor deletion applying. This form is derived not via the application of phonological processes, but rather through the explicit coding of the output as its UR. Alternatively, this form could have been generated from an input /imatʃ-_-i/ with a general depalatalization process that transforms [tʃ] to [t] across all contexts. Under this hypothesis, the output is formed via the application of a phonological process not a part of the original grammar used to generate the data. In other words, the output is compatible with a number of different hypotheses, many of which are completely disjoint from the grammar assumed to be learned. As a result, it is hard to interpret which grammars are learned. In the following sections, we present an alternative account of why learners perform better in some languages and certain forms in those languages: they have more distinct grammars that can generate them.

Our argument is similar to that put forth by Rafferty et al. (2013), who argued that another assumption made by experimentalists and computational modellers – the link between learnability and typological frequency – is incomplete. They showed computationally that the number of hypotheses that can generate a particular language, weighted by their prior, is an indicator of both its typological frequency, as well as its *stability over time*. This line of reasoning is also echoed in Martin and White (2021) who suggest that asymmetries in the learnability of harmony and disharmony pattern may emerge due to differences in the number of compatible analyses for each pattern. The results here are parallel to this argument.

The idea that learners' behavior reflects the combined influence of many different hypotheses about the grammar is consistent with a broad class of models known as *noisy channel models* (Feldman & Griffiths, 2007; Levy, 2008), which

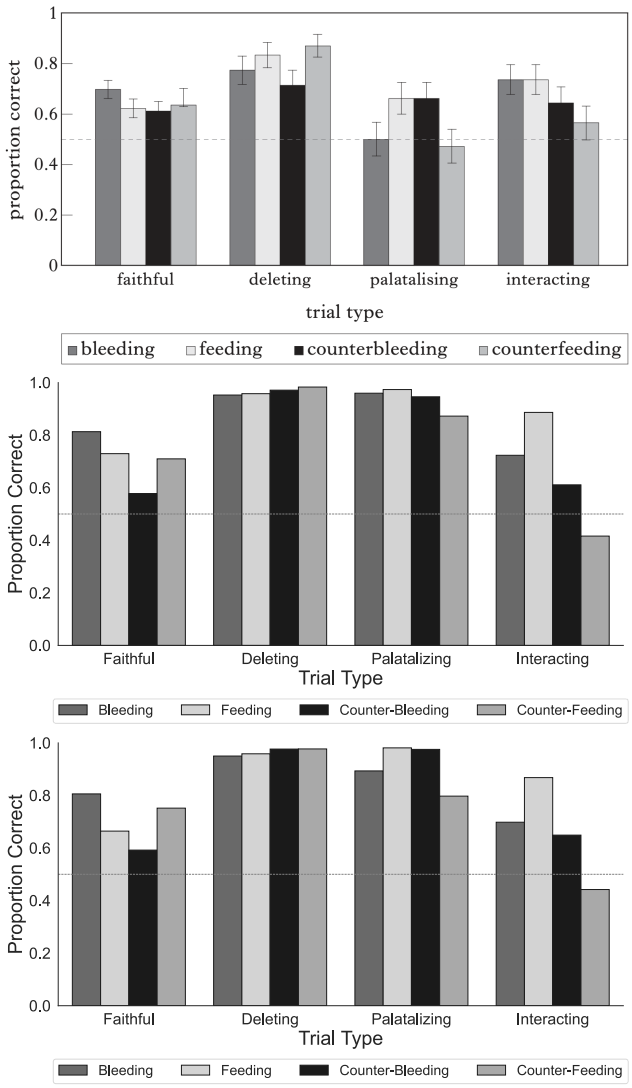

Figure 1: **Top**. Average proportion of correct answers in the testing phase. Plot taken from Prickett (2019). **Middle.** Predictions under a model biased towards only consistent grammars. The model makes similar qualitative predictions as the experimental results in all trial types except in the palatalizing trials, where the bleeding language outperforms the counter-bleeding language. **Bottom.** Predictions under a model that considers both consistent and nearly-consistent grammars. The model makes the same qualitative predictions as those seen in the empirical results.

have previously been applied in the acquisition literature to syntactic learning (Perkins, Feldman, & Lidz, 2017; Schneider, Perkins, & Feldman, 2020). Those models assume that the data learners observe may have been corrupted by a noise process, which creates additional uncertainty about what the uncorrupted data look like. This forces them to consider not only hypotheses that are consistent with the data, but also hypotheses that are almost consistent with the data.

Model

Data. We assume that a datum D_x consists of two parts: the sequence of atomic meanings, or lexemes $\mu_{D_x} = \langle m_i, m_j, ... \rangle$ and the surface form $f_{D_x} = \langle s_{m_i}, s_{m_j}, ... \rangle$. A sample dataset given to the model is shown in (4).

				(4)
imat$_{\langle m_1 \rangle}$	imat-a$_{\langle m_1, m_2 \rangle}$	imatʃ-i$_{\langle m_1, m_3 \rangle}$	imatʃ-_-i$_{\langle m_1, m_2, m_3 \rangle}$	
imak$_{\langle m_4 \rangle}$	imak-a$_{\langle m_4, m_2 \rangle}$	imak-i$_{\langle m_4, m_3 \rangle}$	imak-_-i$_{\langle m_4, m_2, m_3 \rangle}$	
ulit$_{\langle m_5 \rangle}$	???$_{\langle m_5, m_2 \rangle}$	???$_{\langle m_5, m_3 \rangle}$	???$_{\langle m_5, m_2, m_3 \rangle}$	
ulik$_{\langle m_6 \rangle}$	???$_{\langle m_6, m_2 \rangle}$	???$_{\langle m_6, m_3 \rangle}$	???$_{\langle m_6, m_2, m_3 \rangle}$	

We show the model data like in (4), then ask it to generate the forms it has not seen before (those marked with '???').

Setup. We employ a noisy channel model of morphophonological learning. The model generates an output in two steps. In the first step, a grammar, composed of a set of underlying forms U and sequence of mappings R, generates a set of intended outputs I. In the second step, noise potentially corrupts the intended outputs, generating the observed forms D. U and R are both categorical variables. A grammar then consists of a set of fully specified underlying forms for all valid lexeme combinations.

$$D = \langle \langle f_1, m_1 \rangle, \langle f_2, m_2 \rangle, ..., \langle f_n, m_n \rangle \rangle \quad (5)$$
$$U = \langle \langle u_1, m_1 \rangle, \langle u_2, m_2 \rangle, ..., \langle u_n, m_n \rangle \rangle$$

For each individual form, we assume a single unique UR. Lexemes will always take the same UR for a given context, but can vary across different contexts. For example, for the forms [imat] $\langle m_1 \rangle$ and [imatʃ-i] $\langle m_1, m_2 \rangle$, a possible UR hypothesis could consist of /imat/ and /imatʃ-i/. While m_1 shows up in both forms, it does not have the same UR in both. [t]-final stems can either have a UR that is [t]-final or [tʃ]-final. The suffixes can either have a faithful UR or a null UR. The set of underlying forms for the toy languages are given in Table 3. We assume a uniform prior over URs.

In addition to the deletion and palatalization processes discussed in (3), we posit the following rules:

Depalatalization: tʃ → t / _# (6)
Generalized depalatalization: tʃ → t
Generalized palatalization: t → tʃ

The first rule states that the palatalized consonant [tʃ] depalatalizes to [t] when found at the end of the word. The second and third rules are generalizations of the depalatalization and palatalization processes, which apply the relevant process across all contexts. These rules were incorporated to allow the grammar to generate forms such as the faithful form [imat] from URs that contain the palatalized consonant (i.e. /imatʃ/). These rules were also added as doing so ensured that there was no *a priori* asymmetry in the ability to generate [t] vs. [tʃ]. Any asymmetry in performance will arise as a consequence of the space of grammars with respect to the data rather than an arbitrary choice in the hypothesis space. The full space of individual rules is given in Table 3.

A rule hypothesis R is generated by taking some ordered subset over the space of individual rules. For example, one

Table 3: Hypothesis space used in the simulation for the toy languages. The first column corresponds to the space of individual rules. The following columns correspond to example UR hypotheses for the (counter-)bleeding and (counter-)feeding languages.

	(Counter-)bleeding			**(Counter-)feeding**		
Rules	*Faithful*	*Palatalizing*	*Interacting*	*Faithful*	*Palatalizing*	*Interacting*
Palatalization	/imat-a/	/imat-i/	/imat-i-a/	/imat-a/	/imat-i/	/imat-a-i/
Depalatalization	/imatʃ-a/	/imatʃ-i/	/imat-_-a/	/imatʃ-a/	/imatʃ-i/	/imat-_-i/
Generalized Palatalization			/imatʃ-i-a/			/imatʃ-a-i/
Generalized Depalatalization			/imatʃ-_-a/			/imatʃ-_-i/
Deletion	–	–	*Deleting*	–	–	*Deleting*
	–	–	/imak-i-a/	–	–	/imak-a-i/
	–	–	/imak-_-a/	–	–	/imak-_-i/

possible rule hypothesis is to have deletion followed by generalized palatalization followed by depalatalization. We stipulate a fixed meta-ordering among rules: more specific rules must precede more general rules (see Kiparsky, 1973 for discussion and reasoning why). As was the case with the URs, we assume a uniform prior over rules.

We assign the probability of generating an intended output I given the rules and URs as a 1-0 probability:

$$P(I_x|U_x,R) = \mathbb{1}\left[R(U_x) = I_x\right] \tag{7}$$

In the second step, the intended output is then potentially corrupted by noise. We adopt the framework used in Levy (2008) and formalize this noise by computing the Levenshtein distance $L(\cdot)$ (Levenshtein et al., 1966) between the intended and observed outputs. Surface forms with greater Levenshtein distance to the intended output are less probable:

$$P(D_x|I_x) \propto e^{-\lambda L(I_x \to D_x)} \tag{8}$$

The hyperparameter λ controls how much noise is tolerated; the higher λ, the less noisy the output.

The likelihood of a dataset is the product of the likelihoods of each individual datum.

$$P(D|I) = \prod_{x}^{|D|} P(D_x|I_x) \tag{9}$$

Inference and generating predictions. Ultimately, we want the model to produce forms it has both seen and not seen before. To do this, we compute the posterior over grammars $P(U,R|G)$ using Bayes' rule. Since we are operating with categorical grammars, there is only one intended output for each grammar: the output of the grammar $R(U)$. Thus, the likelihood is straightforward to compute.

$$P(U,R|D) \propto P(D|R(U))P(U,R) \tag{10}$$

We estimate the posterior distribution via MCMC using Gibbs sampling (Geman & Geman, 1984). We train the model for 500,000 iterations, discarding the first half and sampling every 25 samples. We generate predictions via the posterior predictive distribution:

$$\begin{aligned} P(d^*|D) &= \mathbb{E}_{P(U,R|D)}\left[P(d^*|U,R)\right] \\ &\approx \frac{1}{N}\sum_{i}^{N} P(d^*|U_i,R_i), \quad U_i,R_i \sim p(U,R|D) \end{aligned} \tag{11}$$

Accuracy is computed as in Prickett (2019), where we renormalize the probabilities of each option:

$$Accuracy = \frac{P(intended|D)}{P(alternative|D) + P(intended|D)} \tag{12}$$

Results

We ran the model twice using different values of λ. In one run, we set $\lambda = 100$ in order to examine performance under a model that only considered consistent grammars. In the other run, we set $\lambda = 3$ in order to evaluate performance under a model that considers both consistent and nearly-consistent grammars. We assess the models' performance by examining their accuracies for each language on held-out data. The results are given in Figure 1. Under a model biased towards only consistent grammars, we observe higher performance for the feeding and bleeding languages in both the palatalizing and interacting trials, contrasting what was observed empirically. Under a model that entertains both consistent and nearly-consistent grammars, we find that the model succeeds in capturing both of the empirical observations made by Prickett. To illustrate how the model achieves the results, we first discuss the effect of consistent grammars on performance before moving on to the effect of nearly-consistent grammars.

Consistent grammars. The expected interacting forms in the feeding and bleeding languages have more high-posterior grammars tied to them (i.e. occupy a larger share of the grammatical space) as the grammatical requirements needed to generate them are less restrictive than the counter-feeding and counter-bleeding forms. The feeding language allows the model to infer grammars that posit the necessary palatalization process for all of its UR hypotheses, whereas the counter-feeding language does not. For example, for the feeding form [imatʃ-_-i], the model can jointly posit a UR like /imatʃ-_-i/ and a vacuous palatalizing process as the two do not compete. In contrast, for the counter-feeding form [imat-_-i], the model cannot posit a palatalization process if its corresponding UR hypothesis is /imat-_-i/. Having fewer compatible UR-rule combinations results in only a fraction of hypotheses acquired in the learning process actually being able to generate the correct output, ultimately lowering accuracy. The

same also holds for the bleeding and counter-bleeding languages. Each UR hypothesis for the bleeding output is compatible with a number of different rule hypotheses while still being consistent with the other forms of the language. For example, [imat-_-a] can be generated via the UR /imatʃ-_-a/ and a rule hypothesis containing the generalized de-palatalization process. This requirement still allows us to posit other rule hypotheses to generate the other observed forms of the language (e.g. palatalization to capture [imatʃ-i]). In contrast, the counter-bleeding output can only be generated by a subset of the UR hypotheses; no rule hypothesis can generate all the forms of the counter-bleeding language if the UR of the interacting form is /imat-_-a/. Like in the case of the feeding vs. counter-feeding languages, this greatly restricts the space of possible ways of generating the counter-bleeding output, lowering performance.

The palatalizing trials have the same UR hypotheses across all four languages, so variation in performance on those trials across languages must stem from differences in the training data: specifically, the interacting forms. A consequence of learning from interacting forms is that each language has a different number of grammars that include the necessary palatalization process. We find indeed that the number of palatalizing grammars for each language correlates with the model's accuracy: 93% and 95% of the unique grammars sampled by the models trained on the bleeding and feeding languages had a productive palatalization process compared to only 89% and 74% of grammars for the counter-bleeding and counter-feeding languages.

Nearly-consistent grammars. Under a particular hypothesis space, some languages have more closely similar languages in the space than others. As the only difference in performance between this model and the previous model occurs with respect to the bleeding and counter-bleeding languages, we focus our attention there. Consider the difference in the distribution of words in each language:

Bleeding	imat	imat-a	imatʃ-i	imat-_-a	(13)
Counter-bleeding	imat	imat-a	imatʃ-i	imatʃ-_-a	

The bleeding language has more words containing [imat] than [imatʃ]. In contrast, the counter-bleeding language consists of an equal number of [imat] and [imatʃ] forms. Recall that the noisy channel allows the model to consider grammars *almost* consistent with the data. A reasonable alternative grammar for the bleeding language would be to eliminate [tʃ] from the surface entirely (i.e. depalatalize across the board). Grammars of this type have a higher relative posterior probability when trained on the bleeding language than the counter-bleeding language; eliminating [tʃ] would incur only one error in the case of the bleeding language, but two in the case of the counter-bleeding language. Being more likely to adopt grammars that try to produce [t] across all contexts would improve performance on the interacting trials, but worsen performance on the palatalizing trials. The contribution of the space of nearly-consistent grammars generates an effect sufficient to alter the asymmetry observed in the palatalizing trials from favoring the bleeding and feeding languages to favoring the feeding and counter-bleeding languages. It is crucially through the interaction of the space of consistent and nearly-consistent grammars that we achieve the same qualitative results as those seen experimentally.

Discussion

A basic assumption in experimental and computational work is that production results are indicative of how easy a phonological phenomenon is to learn. In this paper, we presented an alternative explanation: certain patterns are easier to produce because they have many grammars that could have generated them. We tested this hypothesis by implementing a noisy channel morpho-phonological learner, which makes predictions based on a space of grammars. We found that the model is able to produce qualitatively similar predictions to what was seen experimentally. This was achieved not only because certain patterns have more grammars that perfectly match the data, but moreover because some patterns have more grammars that *nearly* match the data. These simulations illustrate that the experimental results can be modelled not as a result of successfully learning the intended process interaction, but through the combination of landing on alternative consistent analyses as well as mislearning the data.

There are several avenues of future research. First, while our work employs a rule-based formalism, other approaches, such as Optimality Theory (Prince & Smolensky, 2004) and MaxEnt grammars (Goldwater & Johnson, 2003), rely on constraint interaction. Each theory has different hypothesis spaces and may have different predictions about learnability. It is worth exploring what these theories predict about the distribution of outcomes. Moreover, artificial grammar learning is a widely utilized behavioral paradigm used to assess learnability. Other experiments assessing the learnability of process interactions have likewise been explored (Ettlinger, 2008; Kim, 2012; Brooks et al., 2013). Whether grammatical spaces can capture these asymmetries as well remains an open question. Lastly, this model generates these results under a uniform prior over both rules and underlying forms. Many phonological theories have proposed several substantive biases in order to capture asymmetries in the typological frequencies of different phonological phenomena. These primarily pertain to biases over possible mappings (Smolensky, 1996; Steriade, 2001), but some intuitions on possible UR hypotheses (i.e. having as few URs as possible for a given lexeme) have also been assumed in most of the modelling literature. It would be interesting to explore how different biases and prior will interact with the grammatical space.

Broadly, our work contributes to the phonological learning literature by computationally investigating a basic assumption made by the field. We encourage future work to consider the role of the *space* of grammars when interpreting production results.

References

Brooks, K. M., Pajak, B., & Baković, E. (2013). Learning biases for phonological interactions. In *Poster presented at 2013 meeting on phonology.*

Casali, R. F. (1997). Vowel elision in hiatus contexts: Which vowel goes? *Language*, 493–533.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. MIT Press.

Ettlinger, M. (2008). *Input-driven opacity*. University of California, Berkeley.

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory* (pp. 111–120).

Kim, Y. J. (2012). Do learners prefer transparent rule ordering? An artificial language learning study. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 48). Chicago, Illinois.

Kiparsky, P. (1968). Linguistic Universals and Linguistic Change. In E. Bach & R. T. Harms (Eds.), *Universals in linguistic theory.* New York: Holt, Rinehart, and Winston.

Kiparsky, P. (1971). Historical Linguistics. In W. O. Dingwall (Ed.), *A survey of linguistic science* (pp. 576–642). College Park, MD: University of Maryland Linguistics Program.

Kiparsky, P. (1973). 'Elsewhere' in Phonology. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle.* New York: Holt, Rinehart and Winston.

Levenshtein, V. I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 234–243).

Martin, A., & White, J. (2021). Vowel harmony and disharmony are not equivalent in learning. *Linguistic Inquiry*, *52*(1), 227–239.

Perkins, L., Feldman, N., & Lidz, J. (2017). Learning an input filter for argument structure acquisition. In *Proceedings of the 7th workshop on cognitive modeling and computational linguistics (cmcl 2017)* (pp. 11–19).

Prickett, B. (2019, nov). Learning biases in opaque interactions. *Phonology*, *36*(4), 627–653.

Prince, A., & Smolensky, P. (2004). Optimality theory: Constraint interaction in generative grammar. In J. McCarthy (Ed.), *Optimality Theory in Phonology* (pp. 1–71). John Wiley & Sons, Ltd.

Rafferty, A. N., Griffiths, T. L., & Ettlinger, M. (2013). Greater learnability is not sufficient to produce cultural universals. *Cognition*, *129*(1), 70–87.

Schneider, J., Perkins, L., & Feldman, N. H. (2020). A noisy channel model for systematizing unpredictable input variation. In *Proceedings of the 44th annual boston university conference on language development* (pp. 533–547).

Smolensky, P. (1996). The initial state and 'Richness of the Base' in Optimality Theory. *Rutgers Optimality Archive*, *293*.

Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. *Ms., UCLA*.