Phonological Interactions, Process Types, and Minimum Description Length Principles

Christopher Yang
Department of Linguistics and Philosophy
MIT
cminwoo@mit.edu

Kevin Ellis
Department of Computer Science
Cornell University
kellis@cornell.edu

Abstract

Learnability has been a topic of great interest in phonology. In particular is the question of the relative learnability of process interactions. In both historical and experimental domains, researchers have noted that certain kinds of interactions are harder to learn than others. In both domains, however, the results are seemingly in conflict. One potential source of the conflicting outcome is the types of processes involved. In this paper, we investigate the effect of process types on the learnability of different interaction types, using an ideal minimum-description-length learner (MDL). We find that the model indeed predicts different learnability outcomes for each interaction type; however, the asymmetry is largely independent of the process type. This computational model explains certain elements of both the historical as well as some experimental findings of the relative learnability of linguistics process interactions, while contradicting other behavioral findings.

Keywords: Phonology, Bayesian, Linguistics

Introduction

The puzzle of understanding how humans acquire language is a long-standing challenge within the linguistic and cognitive sciences. Human language learning is marked by both its *depth* and its *speed*: somehow, the child masters the entire linguistic stack, from phonetics to semantics, all in less than a decade. A long tradition within linguistics has focused on partitioning language skills into different modules (morphology, pragmatics, etc.), as this piecemeal approach makes the problem more tractable (cf. Modularity of Mind (Fodor, 1983)). A complementary tradition within the cognitive sciences is to carefully engineer *artificial grammars* and study how humans acquire language patterns within these carefully controlled micro-languages.

Here we computationally study how different grammatical processes interact. We construct a space of artificial grammars which probe linguistic phenomena within the domain of *phonology*, the system which explores how sounds are categorized, how sounds change based on their context via phonological processes, and how the these processes interact with one another. There have been several behavioral studies that have looked at the difference in the learnability of different process interactions (Ettlinger, 2008; Kim, 2012; Brooks, Pajak, & Bakovic, 2013; Prickett, 2019). Yet the direction of the asymmetric learnability varies across experiments.

We thus seek to answer the following question: what properties of certain interactions cause them to be more or less learnable, and what properties do not? We make the following contributions:

- We compare the learnability profiles of pairs of nearly-identical languages that differ only in the individual processes involved. Our "map" of the space of process interactions spans existing behavioral studies while suggesting new as-yet unexplored studies, which we hope spurs further investigation.
- We construct an ideal minimum-description-length learner (MDL: (Solomonoff, 1964; Rissanen, 1978; Rasin, Berger, Lan, & Katzir, 2018)) for rule-based phonology (Chomsky & Halle, 1968). This model makes predictions about human linguistic generalization within this space of interactions.
- We find that the model indeed aligns with certain principles of historical change and some behavioral evidence, but contradicts others. Crucially, we find that the model is indeed sensitive to the process types involved, but not in a fashion that is compatible with the empirical evidence.

We begin first with some background on phonological processes and their interactions. We then go over the empirical data, as well as the artificial languages we will be using to pursue the question laid out above. We will then present the model, and conclude with the results of the simulations and a comparison to the empirical data.

Linguistic background

One of the goals in generative phonology is characterizing how infants acquire a phonological grammar from a set of surface forms (e.g. words). A phonological grammar is comprised of two essential parts: the mental representations (also known as the underlying representation, or UR) and the mapping from URs to observed surface strings. For example, many languages insert a vowel between two consonants: /batn/ $\rightarrow$ [batin][1]. Some theories (Chomsky & Halle, 1968)

[1]Slashes '/' correspond to input strings, whereas square brackets '[]' correspond to output strings (so-called surface strings). The arrow '$\rightarrow$' simply denotes that some input string becomes the output string. If some phonological process applies, then the input and outputs look different; if no process applies or if the process applies vacuously, then the input and output are identical.

represent these mappings as rules that sequentially apply to transform the UR to its observed output:

$$X \rightarrow Y \ / \ A\_B \tag{1}$$

A rule of the form as in (1) looks for instances of AXB, also known as the structural description. If such an instance is found, the rule transforms X into Y. For example, for an insertion process, we have[2]

$$\varnothing \rightarrow i \ / \ C\_C \tag{2}$$

For /batn/, the rule looks through the string and locates an instance in which the structural description is met - in this case, sequences of two consonants: C∅C or CC - and inserts [i] in the context. Since the bigram [tn] satisfies the structural description of the rule, [i] is inserted, resulting in [batin].

While (2) corresponds to a single phonological process, individual processes may interact with one another. For example, in addition to insertion, a language may have a process of palatalization, where [t] becomes [tʃ] before [i]: t → tʃ / _i e.g. /bati/ → [batʃi]. Given an underlying string /batn/, if we first apply insertion, we generate intermediate [batin]. The insertion of [i] creates an environment in which palatalization can then apply, generating [batʃin]. For the remainder of the paper, we focus on two-process interactions; or, interactions involving only two independent phonological processes.

There are two dimensions to which these interactions are characterized: whether process A creates or eliminates the environment for which process B applies, and whether process A precedes or follows process B. For example, consider the following rules:

$$\begin{aligned} &\textbf{A}: W \rightarrow Y \ / \ \_X \\ &\textbf{B}: Y \rightarrow Z \end{aligned} \tag{3}$$

The application of A in (3) *generates* the environment to which B can then potentially apply. If A precedes B, A can successfully create the environment for B. This is known as a *feeding* interaction. In contrast, if B precedes A, then B will not apply, as A occurred too late in the phonological interaction. This is known as a *counter-feeding* interaction.

Consider now the following set of rules:

$$\begin{aligned} &\textbf{A}: X \rightarrow Y \\ &\textbf{B}: W \rightarrow Z \ / \ X\_ \end{aligned} \tag{4}$$

Now, the application of process A in (4) eliminates the environment for which B can apply; in A, transforming X to Y necessarily prevents B from applying. If A precedes B, A can successfully destroy the environment for which B would have applied. This is known as a *bleeding* interaction. In contrast, if B precedes A, then A cannot destroy the environment for B. This is known as a *counter-bleeding* interaction.

[2]C and V correspond to consonants and (vowels, respectively. The empty set ∅ corresponds to nothing (i.e. no sound).

Table 1: Experimental results by process type combination

	Experiment	Results
Identity-Identity	Ettlinger 2008	{CF, CB} > {F, B}
Deletion-Identity	Prickett 2019	F > {B, CB} > CF

Process types, maximal utilization, and opacity. There are roughly three types of phonological processes: identity, deletion, and insertion processes. Identity processes involve changing an underlying feature [F] of a sound from one value [αF] to another [−αF] (e.g. the voicing feature of $[p]_{[-voice]} \rightarrow [b]_{+voice}$). Deletion processes involve deleting individual phonemes (e.g. the deletion of [a] in /batai/ → [bati]). Finally, insertion processes insert segments (e.g. the insertion of [i] in /batn/ → [batin]. This characterization is given in (5) below.

$$\begin{aligned} &\textbf{Identity}: [\alpha F] \rightarrow [-\alpha F] \\ &\textbf{Deletion}: X \rightarrow \varnothing \\ &\textbf{Insertion}: \varnothing \rightarrow Y \end{aligned} \tag{5}$$

Researchers have noticed that different process interactions are preferred over others (i.e. are more commonly found in the world's languages or easier to learn); in particular, within the domains of historical change (Kiparsky, 1968, 1971) and learnability (Ettlinger, 2008; Brooks et al., 2013; Kim, 2012; Prickett, 2019). There have been several attempts at generalizing this difference; in particular, that of the maximal utilization (Kiparsky, 1968), and transparency (Kiparsky, 1971) biases.

The maximal utilization bias stipulates that orders that maximize the application of rules (i.e. feeding and counter-bleeding) are preferred over orders in which fewer rules apply (i.e. bleeding and counter-feeding). For example, in the case of feeding such as in (3), both A and B sequentially applied to produce the output. In contrast, for the bleeding order in (4), only A applies, as the application of A destroyed the environment for B to apply.

The transparency bias partitions the space differently, with interactions assigned as either transparent or opaque. Opaque interactions correspond to one of two conditions: either a sound that should have changed in some environment does not on the surface, or a sound that should not have changed because the environment was not met did change. The counter-feeding and counter-bleeding interactions satisfy each of the conditions laid out above, respectively. For example, consider the rules in (3), and an input such as /WX/. If B applies first, followed by A, we would get the output [YX]. On the surface, even though B's structural description is met, the process does not apply. This is contrast to both the feeding and bleeding orders, in which we can observe the respective environments in which each individual process did or did not apply. For example, given the same rules and input as above, if A applies first, followed by B, we produce the output [ZX]. We see the environments for which both A and B

Table 2: Artificial language data

	Feeding	Bleeding	Counter-feeding	Counter-bleeding
Ident-Ident	**A**: e → i / _i **B**: t → d / _i /bat-e-i/ →$_A$ [bat-i-i] →$_B$ [bad-i-i]	**A**: i → e / _e **B**: t → d / _i /bat-i-e/ →$_A$ [bat-e-e] →$_B$ [bat-e-e]	**B**: t → d / _i **A**: e → i / _i /bat-e-i/ →$_B$ [bat-e-i] →$_A$ [bat-i-i]	**B**: t → d / _i **A**: i → e / _e /bat-i-e/ →$_B$ [bad-i-e] →$_A$ [bad-e-e]
Del-Ident	**A**: V → ∅ / _V **B**: t → d / _i /bat-a-i/ →$_A$ [bat-_-i] →$_B$ [bad-_-i]	**A**: V → ∅ / _V **B**: t → d / _i /bat-i-a/ →$_A$ [bat-_-a] →$_B$ [bat-_-a]	**B**: t → d / _i **A**: V → ∅ / _V /bat-a-i/ →$_B$ [bat-a-i] →$_A$ [bat-_-i]	**B**: t → d / _i **A**: V → ∅ / _V /bat-i-a/ →$_B$ [bad-i-a] →$_A$ [bad-_-a]

applied: _(X). The bias stipulates preference for transparent over opaque interactions.

When examining the empirical data, we observe support for both both biases. In the domain of historical change, there have been both observations of re-orderings from *bleeding* by speakers of one generation to *counter-bleeding* in the next, supporting the maximum utilization bias (Kiparsky, 1968) but also cases in the exact opposite direction, with re-orderings from *counter-bleeding* to *bleeding*, supporting the transparency bias (Kiparsky, 1971). Experimentally, participants have exhibited different learnability patterns, even when examining the same interaction types (e.g. feeding vs. counter-feeding).

In Table 1, we observe the outcomes of two different artifical language learning studies, each producing seemingly contradictory results. In Ettlinger (2008), it was found that opaque interactions were easier to learn than transparent interactions, contrary to both the maximal utilization and transparency biases. In contrast, Prickett (2019) found evidence of both the maximal utilization and transparency bias, conditioned on what surface form the grammar must produce.

In the historical and experimental data, we observe differences in the relative preference and learnability of process interactions, with some evidence pointing to a maximal utilization bias, some pointing to an transparency bias, and some pointing to neither. However, in both, this difference was conditioned on another factor: the types of processes involved in the interaction. Historically, the maximal utilization bias was supported primarily by interactions featuring identity-identity languages, whereas the transparency bias was supported primarily by interactions featuring deletion-identity and insertion-identity languages. Experimentally, the relative differences in learnability - although not falling neatly into a pattern predicted by the maximal utilization or transparency bias - also appear to cross-cut this difference. Thus, we seek not only to determine whether an asymmetry between process interactions is predicted by an ideal minimum-description-length learner, but moreover examine how different processes involved in those interactions (e.g. deletion-identity vs. identity-identity) influence that asymmetry.

Artificial language data. Two artificial languages inspired from the artificial data used by Prickett (2019) were constructed. They comprise the identity-identity languages and deletion-identity languages. This is shown in Table 2. The identity-identity languages are composed of a raising process, in which [e] raises to [i] before another [i], and a voicing process, in which [t] becomes [d] before [i][3]. The deletion-identity languages are comprised of a deletion process, in which a vowel deletes before another vowel, as well as the same voicing process as the identity-identity languages. The languages were constructed to minimally deviate from each other, except in terms of the process type involved. This will allow us to detect whether process type affects the model predictions.

The data was organized into paradigms of surface, pronounced speech. A paradigm consists of the root (e.g. "walk") in isolation, as well as conjugated forms (e.g. "walk", "walked", etc.), formed from the concatenation of the root and one or more suffixes (e.g. "walk-ed"). Roots (RT) were of the form $\{b,p,k,g\}\{a,i\}\{t,d,k,g\}$, with sounds equally distributed for each position. Two affixes were assigned for each language: /-e/ and /-i/ for the identity-identity language, and /-a/ and /-i/ for the deletion-identity language. Sample data of the feeding language for each process type is given below[4].

Identity-Identity

RT	RT-e	RT-i	RT-e-i
bat	bate	badi	badii
pit	pite	pidi	pidii
gad	gade	gadi	gadii
kik	kike	kiki	kikii

Deletion-Identity

RT	RT-a	RT-i	RT-a-i
bat	bata	badi	badi
pit	pita	pidi	pidi
gad	gada	gadi	gadi
kik	kika	kiki	kiki

[3]This specific process is a rare if not unattested phonological pattern in language. This process *is*, however, isomorphic to other attested phenomenon, such as palatalization, i.e. t → tʃ / _i.

[4]Note that the artificial language used in (Ettlinger, 2008) differs from the one used in our simulation. Ettlinger's artificial language consists of the following processes: a plural suffix /-il/ surfaces as [-el] when when the preceding vowel is [e], and [e] lowers to [a] when the preceding vowel is [a].

$$\begin{aligned} &\textbf{A}\text{: -il} \rightarrow \text{-el} \;/\; \text{e}C_0\_ \\ &\textbf{B}\text{: e} \rightarrow \text{a} \;/\; \text{a}C_0\_ \end{aligned} \tag{6}$$

Process A in (6) is unlike those described before in the paper, targeting the sound only in a specific suffix rather than across the entire language. The interaction being observed does not appear to be a phonological one, but a more complex interaction between two modules of the grammar - particularly, the morphology and the phonology. This kind of interaction cannot be stated in purely phonological terms, and such may produce results different from the results achieved in the model.

Model

Setup. In formal terms, our model observes (takes as input) a set X of pronounced surface forms paired with their constituent atomic meanings. For example, if the model observes the English past tense word "walked", one such tuple would be $\langle /\text{wɔkt}/, \textsc{Walk} + \textsc{Past}\rangle$, where WALK and PAST are atomic meanings for "walk" and the past tense, respectively. It infers (i.e. outputs) a sequence of K ordered rules, written $\{r_k\}_{k=1}^{K}$, as well as a *lexicon*. The lexicon, written $\mathbb{L}$, is a function from a meaning atom to a sequence of phonemes (its underlying representation). Together the rules and lexicon make a *grammar*. Given this notation, the constraint C that the inferred grammar satisfies a particular example $\langle f, m_1 + m_2 + \cdots m_N\rangle$ is

$$C(\langle f, m_1 + m_2 + \cdots m_N\rangle, \{r_k\}, \mathbb{L}) = \mathbb{1}\left[f = \textsc{Phonology}(\mathbb{L}(m_1)\cdot\mathbb{L}(m_2)\cdots\mathbb{L}(m_N))\right] \quad (7)$$

$$\text{where: } \textsc{Phonology}(f) = r_1(r_2(r_3(\cdots f))) \quad (8)$$

While we generally seek grammars consistent with the input data, this desiderata trades off against a bias for a parsimonious grammar. One way of encoding this trade-off is through so-called minimum-description-length (MDL) models. MDL is a method for finding a single model given a dataset, which seeks to minimize the sum of the size of the model and the size of the data given the model. We follow the classic scheme of identifying the model with the grammar, and the size of the data given the model as the size of the lexicon needed given those rules (Ellis, Solar-Lezama, & Tenenbaum, 2015; Rasin et al., 2018; Barke, Kunkel, Polikarpova, Meinhardt, et al., 2019). Because the lexicon can memorize any surface forms that do not obey the rules, size of lexicon measures compression of the data given the model.

As an MDL learner, our model works by jointly minimizing the total size of rules, lexicon, while also minimizing the size of forms unexplained by this grammar. Thinking of the rules as a model and the (form, meaning) pairs as observed data, we jointly minimize the description-length size of the model and the size of the data conditioned on the model:

$$\textsc{ModelSize}(\{r_k\}) = \sum_k \text{size}(r_k) \quad (9)$$

$$\textsc{DataSize}(X, \{r_k\}, \mathbb{L}) = \sum_{\langle f,m\rangle\in X} \underbrace{\text{size}(f)}_{\text{MDL...}} \underbrace{(\neg C(\langle f,m\rangle, \{r_k\}, \mathbb{L}))}_{\text{...if datum unexplained}} + \sum_{f\in\text{range}(\mathbb{L})} \text{size}(f) \quad (10)$$

Optimization. In general even finding *any* grammar consistent with the data–let alone a parsimonious one–is computationally intractable. We wish to study the behavior of *idealized* MDL learners, independent of computational implementation, thereby teasing apart the learning objectives from process-level/performance-level (Chomsky, 2014) details. Therefore we adopt a sound and complete optimization procedure using constraint-based program synthesis (Solar-Lezama, 2008), which uses a Satisfiability Modulo Theories (SMT (Barrett, Stump, & Tinelli, 2010)) solver. This computational technique guarantees exact solving of combinatorial objectives such as those above, while sacrificing guarantees on runtime. Concretely we use the Sketch program synthesis engine (Solar-Lezama, 2008), which was also used by (Zuidema et al., 2020) for inducing phonological rules. Sketch can be used to compute Pareto optimal[5] solutions to the joint minimization of MODELSIZE and DATASIZE using an SMT-enabled optimization algorithm. Note that unlike other works examining Pareto optimal program synthesis outputs (Schmidt & Lipson, 2009), our approach yields the exact Pareto frontier, rather than post-hoc calculating the frontier over the outputs of a heuristic search.

Generating predictions. Ultimately we are concerned not with specific grammars output by the model, but rather with what predictions those grammars make on unseen words. After all, behavioral studies of language learning directly yield statistics of human generalization, *not* rules and lexica.

To assess models and human behavior side-by-side we ask the model to predict the pronounced forms of unseen test words. Because the model may discover many Pareto optimal grammars, we accomplish this through Bayesian model averaging: for each grammar, we interpret its description length as a negative log posterior probability, and integrate over the space of Pareto optimal grammars:[6]

$$P(\{r_k\}, \mathbb{L}|X) \propto \mathbb{1}\left[\langle\{r_k\}, \mathbb{L}\rangle \text{ Pareto optimal for } X\right] \times P(X|\{r_k\}, \mathbb{L})P(\mathbb{L})\prod_k P(r_k) \quad (11)$$

$$P(r_k) \propto \exp(-\alpha\,\text{size}(r_k)) \quad (12)$$

$$P(\mathbb{L}) \propto \prod_{f\in\text{range}(\mathbb{L})} \exp(-\beta\,\text{size}(f)) \quad (13)$$

$$P(X|\{r_k\}, \mathbb{L}) \propto \prod_{\langle f,m\rangle\in X} \exp(-\gamma\,\text{size}(f)(\neg C(\langle f,m\rangle, \{r_k\}, \mathbb{L}))) \quad (14)$$

where hyperparameters α, β, γ trade-off between a preference for small rules, small lexicon, and fit to data, respectively.

We are primarily concerned with the description length assigned by the model to held out test data, which is connected to the above equations as follows. Given train/test data X/Y, we marginalize over Pareto optimal models:

$$\text{size}(Y|X) = \sum_{\{r_k\}}\sum_{\mathbb{L}} \textsc{DataSize}(Y, \{r_k\}, \mathbb{L})P(\{r_k\}, \mathbb{L}|X) \quad (15)$$

[5] Formally, pareto optimality for multi-objective optimization means not-strictly-worse than any other solution along all optimization objectives. Intuitively, these are the grammars that a learner might entertain depending on how it relatively weighs parsimony and fit to data.

[6] Ideally one would integrate over the entire infinite space of all grammars, but this is wildly intractable. These equations are written up to a proportionality, which is licensed by the fact that there are finitely many possible grammars/forms, so each probability distribution is normalizable

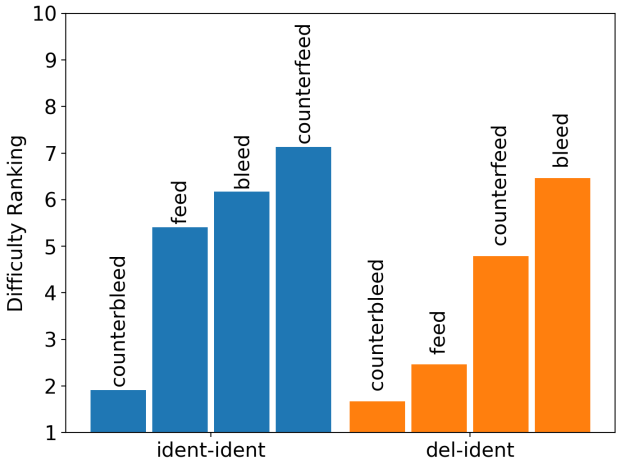

Figure 1: How hard does an ideal MDL learner predict each language should be to acquire? We assess the model on held out test data from each language and score it according to the description length it assigns to this test data. Graphed above are the average ranking of each language according to this difficulty measure. Results averaged over different values of the hyper parameters α, β, γ (Eq. 12-14) uniformly sampled from $[0,2] \times [0,2] \times [1,2]$

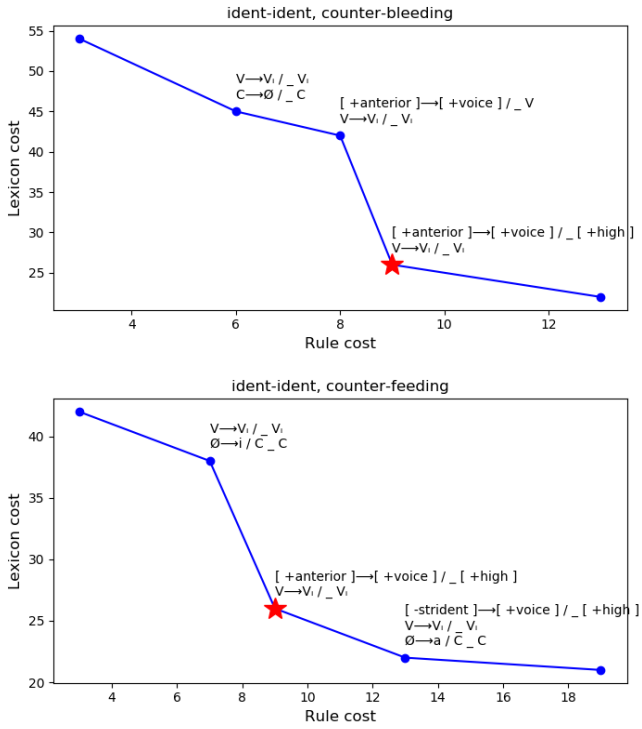

Figure 2: Insight into how the model succeeds with some languages and struggles with other comes from visualization of the space of possible grammars. Here we show Pareto optimal grammars for two languages. Each point is a different grammar. **Top**, easiest language to learn. Around the correct grammar (red) one observes a sharp kink in the Pareto frontier, which is where a Bayesian learner would tend to concentrate its posterior probability. **Bottom** shows an intermediate difficulty language. One observes a less dramatic kink for the medium difficulty. The less prominent kink indicates a broader, more diffuse and uncertain posterior distribution over grammars.

Ultimately, we suggest that an adequate explanation of linguistic generalization competence should assign higher description length to linguistic phenomena which are harder to learn. This is because across many learning strategies, description length should be a good proxy for learning difficulty, e.g. difficulty scales exponentially for enumerative learning strategies, and generally learning is harder for stochastic methods as well. More broadly, description length is a proxy for likelihood under the prior, so higher descriptions length correlates to higher violations of prior expectations.

Results

We want to know whether the model finds it easier to learn the same interactions that behavioral studies suggest humans easily acquire, and vice versa for difficult-to-learn interactions. To answer the question of which interaction types are easier to learn according to the model, we prepare parallel corpora of train and test words from each language. We asked the model to learn grammars from training words and measure the description length it assigns to test words.

Figure 1 plots the difficulty of each process interaction by process type (compare with Figure 2). As the relationship between the opaque and transparent interactions is most clear (i.e. is a result of a simple re-ordering of rules), we focus our comparison on this aspect. For the identity-identity language, feeding and counter-bleeding interactions were found to have a lower description length and thus easier to learn than their counter-feeding and bleeding counterparts. For the deletion-identity language, we again see that feeding and counter-bleeding languages are easier to learn. These results both conform to the maximal utilization bias.

We therefore observe an asymmetry in the learnability of process interactions that implements a maximal utilization bias, *independent* of the process types, and observe no systematic bias toward transparency. We stress that the model holds no explicit knowledge of interaction types as well as process types more broadly; the behavior of the model depends only on the MDL framing, not on its algorithmic implementation.

Why should we observe a maximum utilization bias? First note that this bias says that, if rules apply all the time, then they will be easier to learn. In MDL terms, introducing a new rule "buys" more description length than it costs because the rule is frequently used. Referencing Figure 2 top, there is a sharp drop in the lexicon MDL at the correct grammar, because adding all the correct rules dramatically decreases the

amount of surface forms memorized in the lexicon. From an MDL perspective, we should not expect a transparency bias when it comes at the expense of maximal utilization, independent of process type. While this conforms to earlier analyses of historical change (Kiparsky, 1968) as well as some behavioral studies (Prickett, 2019), the model is unable to utilize the process types involved in the interaction in order to derive the transparency bias (Kiparsky, 1971). This is of potential concern, as the results here seemingly violate more modern analyses of these interaction, which tend to argue in favor of an asymmetry partitioned based on the transparency of the interaction rather than the utilization of the processes.

This suggests that a bias toward transparency, if MDL is to be a viable account of language acquisition for phonology, must come from the algorithmic implementation of linguistic learning mechanisms, not the objective function of grammar learning. In other words, this effect must hinge on linguistic *performance*, instead of linguistic *competence* (Chomsky, 2014). Here we refer to Chomsky's performance-competence distinction, where performance refers to the actual fine-grained behavioral characteristics of language acquisition and use, while competence refers to its abstract characterization in terms of representations and objectives. Absent a process-level account, we propose that MDL learners cannot explain this transparency bias, although they do suffice to explain a maximal utilization bias.

Related work

There are other dominant models of learning phonology, such as Maximum Entropy models (Goldwater & Johnson, 2003) and the Gradual Learning Algorithm (Boersma & Hayes, 2001). Both Maximum Entropy models and the Gradual Learning Algorithm more closely follow the formalism of Optimality Theory (OT; (Smolensky & Prince, 1993)) as opposed to SPE-style rules, as we do here. OT is a more popular and recent alternative to such rule-based phonology, and Bayesian approaches have proved useful here for building computational learning accounts (Goldwater & Johnson, 2003; Doyle, Bicknell, & Levy, 2014). Here we worked out the consequences of MDL learners in rule-based terms, leaving study of other model families to future work.

MDL learners are one of very few models that are able to jointly learn the underlying representations as well as the mapping from the underlying to surface forms. These models have seen a variety of success in modeling different morpho-phonological phenomenon. An earliest application of MDL to morpho-phonological learning is Rasin et al (2018). This is a genetic algorithm that stochastically explores the space of possible rules, and can, among other phenomena, discover opaque interactions. Notably, it exhibits the transparency bias. We believe our findings here suggest that these transparency biases must emerge not from the MDL framing, but from the implementation details of its stochastic search. Complementary work (Barke, Kunkel, Polikarpova, & Bergen, 2019) combines SMT-based methods with bottom-up heuristics and has been applied to learning several different phonological phenomena, such as final devoicing and epenthesis.

We emphasize that we are not concerned here with the exact implementation of MDL learners. Instead we have built a generic MDL model, and exactly solve for its optimal solutions. By disentangling learning objectives from algorithmic implementation, we can make a broader claim about what inductive biases must necessarily emerge from the MDL framing, and what inductive biases must necessarily *not* emerge.

Discussion

Empirical investigations of phonological process learnability had previously identified two contradictory learning biases: (1) a bias toward transparency, and (2) toward maximal utilization. We find that a minimum-description-length learner suffices to capture only one of these learnability asymmetries, and that these results are not based on the process type. Our result is *not* tied to the specific implementation details of our model. Our use of exact program synthesis guarantees exhaustive exploration of the space of grammars; our hyperparameter sweep shows that this result is insensitive to how different description lengths are balanced. This suggests that future work should explore process-level models of phonological performance, and not seek to explain transparency biases in terms of ideal-observer models of linguistic competence.

Narrowly viewed, our work contributes to explaining a puzzling finding in phonological process interactions, suggesting that basic principles of inference such as minimum-description-length are a part of solving this puzzle, but still leave more to explain. Zooming further into phonology, this invites additional questions as to whether different subclasses of opacity, or subclasses of identity/deletion/insertion, are easier/harder to learn. For example, as Table 1 suggests, Insertion-Identity interaction phenomena are underexplored, and indeed we are computationally investigating this at the very moment: "mapping out" the space of interactions, simulating computational learners, and making predictions for new behavioral studies.

Viewed broadly our work contributes to the body of literature showing that rational models can explain human learning patterns (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Yet there are limits to these ideal observer models (Marcus & Davis, 2013). We hope this work helps further delineate their strengths and weaknesses.

Acknowledgments. We would like to thank Adam Albright and Naomi Feldman, as well as the anonymous reviewers for their feedback and suggestions.

References

Barke, S., Kunkel, R., Polikarpova, N., & Bergen, L. (2019). Constraint-based learning of phonological processes. *Empirical Methods in Natural Language Processing*.

Barke, S., Kunkel, R., Polikarpova, N., Meinhardt, E., Bakovic, E., & Bergen, L. (2019). Constraint-based learn-

ing of phonological processes. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 6177–6187).

Barrett, C., Stump, A., & Tinelli, C. (2010). The satisfiability modulo theories library (smt-lib). www. *SMT-LIB. org*.

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, *32*(1), 45–86.

Brooks, K. M., Pajak, B., & Bakovic, E. (2013). *Learning biases for phonological interactions.* Amherst, MA.

Chomsky, N. (2014). *Aspects of the theory of syntax* (Vol. 11). MIT press.

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

Doyle, G., Bicknell, K., & Levy, R. (2014). Nonparametric learning of phonological constraints in optimality theory. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (p. 1094–1103).

Ellis, K., Solar-Lezama, A., & Tenenbaum, J. (2015). Unsupervised learning by program synthesis.

Ettlinger, M. (2008). *Input-driven Opacity*. PhD thesis.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Goldwater, S., & Johnson, M. (2003). Learning ot constraint rankings using a maximum entropy model..

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 *Universals in Linguistic Theory.*

Kiparsky, P. (1971). *Historical Linguistics*.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*.

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

Rasin, E., Berger, I., Lan, N., & Katzir, R. (2018). Learning rule-based morpho-phonology. *Working paper*.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*.

Smolensky, P., & Prince, A. (1993). Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in phonology*, *3*.

Solar-Lezama, A. (2008). *Program synthesis by sketching*. PhD thesis.

Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and control*, *7*(1), 1–22.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., & Gentner, T. Q. (2020). Five ways in which computational modeling can help advance cognitive science: Lessons from artificial grammar learning. *Topics in cognitive science*, *12*(3), 925–941.