Modeling the Acquisition of Phonological Interactions: Biases and Generalization

Brandon Prickett and Gaja Jarosz
University of Massachusetts Amherst

1 Introduction

Phonological processes in a language have the potential to interact with one another in numerous ways (Kiparsky 1968, 1971). Two relatively common interaction types are *bleeding* and *feeding* interactions, with (1) showing hypothetical examples of both (for a similar hypothetical example, see Baković 2011).

(1) *Example of Bleeding and Feeding Interactions*

	Bleeding	Feeding
Underlying Representation (UR)	/esi/	/ise/
[-Low] → [αHigh] / [αHigh]C_ (Vowel Harmony)	ese	isi
[s] → [ʃ] / _[+High] (Palatalization)	-	iʃi
Surface Representation (SR)	[ese]	[iʃi]

In the bleeding interaction above, the underlying form /esi/ undergoes harmony, since the /e/ and /i/ have different values for the feature [High]. Since the harmony is progressive, the /i/ assimilates to the /e/ and can no longer trigger the palatalization process (which occurs *after* harmony), resulting in a surface form of [ese]. The feeding interaction has the opposite effect, with an underlying /e/ becoming a high vowel due to harmony and then triggering the palatalization process to create the SR [iʃi]. The crucial difference between the two interaction types in this example is whether the UR undergoes lowering or raising due to the harmony process.

Two other interactions can be produced using these same URs by applying the two rules in the opposite order. These are called *Counterbleeding* and *Counterfeeding* interactions and are shown in (2).

(2) *Example of Counterbleeding and Counterfeeding Interactions*

	Counterbleeding	Counterfeeding
UR	/esi/	/ise/
Palatalization	eʃi	-
Vowel Harmony	eʃe	isi
SR	[eʃe]	[isi]

In the counterbleeding interaction, both palatalization and harmony are able to change the underlying representation, but the motivation for palatalizing (i.e. a high vowel after an /s/) is erased by the harmony rule. In the counterfeeding interaction, the palatalization rule occurs too early to make any changes to the form, despite vowel harmony creating a potential palatalizing context later on in the derivation. Counterbleeding and counterfeeding interactions are typically called *opaque*, because in the surface forms of the language, certain processes (in this case, palatalization) seem to either over- or under-apply in places where they shouldn't (Kiparsky 1971, McCarthy 1999, Baković 2011).

Questions about the formal representation and acquisition of opaque interactions have figured prominently in the phonological literature for decades. Following the introduction of OT, the debate emphasized the representational choices offered by *parallel* vs. *serial* architectures, but there is also a substantial literature debating the productivity (*grammatical* vs. *exceptional* status) of opaque interactions. In this paper we select four theoretical frameworks that exemplify this range of positions to examine the

[*] The authors would like to thank Eric Baković, as well as the members of UMass Amherst's Sound Workshop and Experimental Lab Group, and the audiences at the USC colloquium and the 17th OCP for their helpful feedback on this project.

Proceedings of *AMP 2020*

predictions these representational options make for learning and generalization. We examine two serial theories that, like the rule-based sketch above, treat all process interactions as grammatical: *Stratal OT* (Kiparsky 2000) and *Harmonic Serialism* (McCarthy 2000). We also examine two parallel theories that differ in their treatment of the processes as grammatical vs. exceptional. *Two-level constraints* (McCarthy 1996) ban certain input-output mappings and provide a way to represent all four process interactions as grammatical in parallel OT. Finally, we also examine a parallel model with *indexed constraints* (Pater 2010) which treat opacity as lexical exceptionality (Sanders 2003).

Recent computational and experimental work has sought to better understand how interacting processes are acquired by both humans (Ettlinger 2008, Prickett 2019) and computational models of phonological learning (Jarosz 2016, Nazarov & Pater 2017, Prickett 2019). However, little work has directly compared the differences in predictions made by various theories of opacity regarding learning and generalization. This is the focus of the present paper. We present computational simulations demonstrating that four constraint-based theories of opacity each make unique predictions about these phenomena.

The paper proceeds as follows: §2 summarizes existing work on both human and machine learning of phonological interactions, §3 presents the analyses for opaque interactions that each of the theories of interest use, §4 describes the novel computational modeling experiments we ran, §5 presents the results of those simulations, and §6 interprets them and discusses their implications.

2 Background

2.1 *Learning biases and phonological interactions* When first discussing the possible ways that phonological processes could interact, Kiparsky (1968, 1971) pointed out that speakers seemed to prefer certain kinds of interactions over others. The preferences Kiparsky observed were based on sound changes where rules were lost or re-ordered diachronically. Kiparsky first proposed a bias preferring interactions that maximally utilize all of the rules involved in a derivation, called *Maximal Utilization Bias* (henceforth MaxUtil). For example, in the feeding and counterbleeding examples above, both harmony and palatalization are applied to forms like /ise/ and /esi/, meaning that a MaxUtil bias would favor such languages. In bleeding and counterfeeding interactions, however, the same URs only undergo harmony – palatalization is bled or counterfed and ends up not applying. The MaxUtil bias disfavors these interactions that provide no evidence for palatalization. Kiparsky (1971) later proposed a bias affecting the acquisition of opaque interactions: a *Transparency Bias*. This bias favors bleeding and feeding over their opaque counterparts, counterbleeding and counterfeeding. The idea behind this bias is that process interactions should be easier to learn if the motivation for their application or nonapplication is visible on the surface.

Jarosz (2016) explored what kinds of learning biases would affect an HS model that was trained on interactions between a vowel-deletion process and a palatalization process. She equipped the model with serial markedness constraints (see §3 for more on these) and gave it three kinds of training data: words where only palatalization was applicable, words where only deletion was applicable, and words that demonstrated an interaction between the two processes (either bleeding, feeding, counterbleeding, or counterfeeding). When the model was trained on data that contained equal proportions of these three types of contexts, there was no strong bias for or against any of the four interaction types. However, when the interacting context was presented substantially more often in the training data compared to the contexts where only the individual processes of palatalization and harmony were applicable, a MaxUtil bias emerged (i.e. the model converged more quickly for the feeding and counterbleeding languages). This occurred because it took longer for the model to learn the palatalization process in this condition since most of the data did not support palatalization. On the other hand, when the interacting context was presented much less frequently, the model learned both individual processes quickly, but took a long time to learn the opaque interaction between them. In this condition, the bleeding and feeding interactions were learned more quickly, showing a Transparency Bias.

Nazarov and Pater (2017) also found transparency bias when training a stratal model on a pattern based on the raising and flapping interactions in Canadian English. When they trained the model on interactions between a vowel raising process and a consonant flapping process, it converged more reliably when the interaction was bleeding than when it was counterbleeding. However, they found that, as they increased the realism of the training data by introducing additional data providing evidence for which stratum each process applied in, the model's transparency bias disappeared.

Artificial language learning has been used to investigate whether Kiparsky's (1968, 1971) biases affect human acquisition. Ettlinger (2008) found an apparent bias against maximally utilizing processes when he tested participants on counterbleeding and counterfeeding interactions. This was a between-subjects design with each participant learning patterns demonstrating only one of the two interactions. When tested on what they'd learned, participants that were assigned to his counterfeeding condition seemed to acquire their pattern better than those that were in his counterbleeding condition—the opposite of what a MaxUtil Bias would predict. However, he didn't train use any transparent languages, so a transparency bias could not be explored.

Prickett (2019) used a similar, between-subjects design, except he tested all four of the main interaction types: bleeding, feeding, counterbleeding and counterfeeding. Following Jarosz's (2016) simulations, the interactions were the result of a vowel deletion and palatalization process. When tested at the end of the experiment using a forced choice task, Prickett's (2019) participants showed no language-wide effect of interaction type. However, when results were broken down by stimulus type, differences across languages became apparent. Participant performance on trials where only palatalization was conditioned (which involved a choice between palatalizing a form or mapping it faithfully) was affected by a MaxUtil Bias. That is, participants in the feeding and counterbleeding conditions had higher accuracy than those in the bleeding and counterfeeding conditions. Additionally, a Transparency Bias (i.e. higher accuracy on bleeding and feeding) emerged in participants' performance when they were given words that demonstrated an interaction between the two rules (which involved a choice between a correct and incorrect ordering of the two processes). In computational simulations, Prickett (2019) found that both Jarosz's (2016) HS model and a neural network captured the kind of biases he observed in humans.

2.2 *Generalizing to novel data when learning phonological interactions* Generalization to novel words has been an important method for investigating the kinds of representations that underlie phonological knowledge (e.g. Halle 1978). This has been extended to artificial language learning experiments, to further probe what kinds of representations and biases influence phonological acquisition in that context (e.g. Finley & Badecker 2009). The existing literature on phonological learning and generalization is extensive, but experimental work focusing on human learning and generalization of opaque and transparent interacting processes is quite limited.

Preceding the Prickett (2019) study described above, Kim (2012) and Brooks et al. (2013) both used artificial language learning experiments to test how participants would generalize from a language that had the potential for interaction. That is, participants saw evidence of two different phonological processes that *could* interact in the appropriate context but were never shown that context in the training phase of the experiment. Then, participants were asked to generalize to a novel interacting word to see whether they would do so in a transparent, opaque, or unexpected way. Kim's (2012) results suggested that participants preferred to generalize in a way that created a counterfeeding interaction, although no inferential statistical tests were conducted on those results. Brooks et al. (2013) found that their participants preferred to not apply either process in interacting contexts, a phenomenon that is likely unattested in natural language.

2.3 *Computational phonology and theories' predictions* One of the ways that computational phonology provides insight into the rest of the field is by showing what theories of phonological representation predict when paired with theories of phonological learning. For example, classic OT (Prince and Smolensky 1993) assumes that any possible ranking of constraints in a particular theory (i.e. that theory's *factorial typology*) predicts a language that should be attested in the real world. However, Staubs (2014) and Stanton (2016) demonstrated that standard, constraint-based theories of stress and footing, when paired with computational modeling, predict systematic learning biases that disfavor languages in the factorial typologies which are un or under-attested (see also Hughto 2018).

While the simulations in this paper do not deal directly with typology, they use a similar approach to explore what theories of phonological representation and learning predict about human behavior in an artificial language learning context. Similar work has been used to test theoretical proposals in the past, such as a bias for natural phonological patterns (Wilson 2006) or surface identity constraints (Gallagher 2013). The simulations described in §4 look at two different predictions made by the theories of interest described in §1: (1) whether each theory has the kind of learning biases that Prickett (2019) observed in his experiment and (2) how each theory generalizes to novel kinds of interactions. While no experimental data yet exists for the latter, we show that each theory makes unique predictions that can be tested in future experiments.

3 Representing phonological interactions

Most constraint-based theories, including those examined in this paper, can capture bleeding and feeding interactions straightforwardly with standard markedness and faithfulness constraints. An example of this using classic OT (Prince & Smolensky 1993) is shown in (3) and (4) for the bleeding and feeding interactions from §1. The constraint definitions we use are: **[si]* assigns one violation for every [s][+High] sequence in the output, *Agree* assigns one violation for every pair of non-low vowels in the output that do not agree in their height, and *Ident(F)* assigns one violation for every segment with a different value for feature *F* in the output and input.

(3) *Constraint-based representation of a bleeding interaction*

/esi/	Agree	*[si]	Ident(Anterior)	Ident(High)
[esi]	W*	W*		L
☞ [ese]				*
[eʃi]	W*		W*	L
[eʃe]			W*	L

(4) *Constraint-based representation of a feeding interaction*

/ise/	Agree	*[si]	Ident(Anterior)	Ident(High)
[ise]	W*		L	L
[isi]		W*	L	*
[iʃe]	W*		*	L
☞ [iʃi]			*	*

In the tableaux above, the same constraint ranking produces both bleeding and feeding interactions, depending on what UR the grammar is given as input. Achieving this requires the markedness constraints motivating vowel harmony and palatalization to be ranked above the faithfulness constraints penalizing the two processes. Some differences across the four theories exist for how these transparent languages can be represented. For example, in HS, multiple steps are required to make both of the changes in the feeding derivation and when using indexed constraints, the bleeding interaction can be represented as an exception to a more general palatalization process (e.g. one that palatalizes [s]'s before any front vowels). However, for the sake of space, the rest of this section will focus on how the theories represent opaque interactions, which is where the main differences between them arise.

Stratal OT (Kiparsky 2000) represents phonological derivations as a series of multiple, independently ranked, OT-style grammars (called *strata*) that a form passes through in a particular order. To capture an opaque mapping, whatever process counterbleeds/counterfeeds the other must occur in an earlier stratum. This is exemplified in the tableaux in (5) and (6), where *Stratum 1* takes the UR of a form as input and passes an intermediate form to *Stratum 2*. Then, Stratum 2 takes that intermediate form as input and outputs an SR.

(5) *Stratal OT representation of a counterbleeding interaction*

Stratum 1: Palatalization

/esi/	Ident(High)	Agree	*[si]	Ident(Anterior)
esi		*	W*	L
ese	W*	L		L
☞ eʃi		*		*
eʃe	W*	L		*

Stratum 2: Harmony

eʃi	Agree	Ident(High)	Ident(Anterior)	*[si]
[eʃi]	W*	L		
[ese]		*	W*	
☞ [eʃe]		*		

In these tableaux, the opaque mappings are achieved by limiting each of the two processes (i.e. palatalization and harmony) to one of the individual strata. Stratum 1 allows palatalization to occur by ranking *[si] over Ident(Anterior), but blocks harmony by ranking Ident(High) over Agree. The second stratum does the opposite, allowing harmony but no palatalization. This produces both counterbleeding and counterfeeding, depending on the UR. While counterfeeding can only be captured if each of the strata are limited to a single process, counterbleeding can also be achieved by a grammar with a Stratum 1 allowing only palatalization and a Stratum 2 allowing both processes. When a counterbleeding form is processed by the second stratum, palatalization would have already applied and applying it again would make no change.

(6) *Stratal OT representation of a counterfeeding interaction*

Stratum 1: Palatalization (inapplicable)

/ise/	Ident(High)	Agree	*[si]	Ident(Anterior)
☞ ise		*		
isi	W*	L	W*	
iʃe		*		W*
iʃi	W*	L		W*

Stratum 2: Harmony

ise	Agree	Ident(High)	Ident(Anterior)	*[si]
[ise]	W*	L		L
☞ [isi]		*		*
[iʃe]	W*	L	W*	L
[iʃi]		*	W*	L

The next theory we consider is HS (McCarthy 2000) with Serial Markedness Reduction (Jarosz 2014). HS forces phonological mappings to occur serially, with each form passing through the same grammar over multiple steps, with at most one change occurring to the form during each of those steps, until it eventually maps faithfully onto itself. Crucially, the HS grammar presented here makes use of *serial markedness (SM) constraints* (Jarosz 2014), which create a limited kind of memory across separate passes through the grammar, allowing it to represent opaque mappings (cf. the "candidate chains" in McCarthy 2007). Each candidate encodes which markedness constraints it has satisfied and in what order (shown in angle brackets), while the SM constraints specify which markedness constraints should be satisfied before others. The SM constraints' take the form $SM(m_1, m_2)$, where m_1 and m_2 are markedness constraints, and violations are assigned if m_2 is satisfied before or simultaneously with m_1. This is demonstrated in (7) and (8) for the opaque interactions. "BOTH" in angle brackets indicates both markedness constraints in the tableau are satisfied in the same step, which incurs a violation for any SM constraint that specifies an order for those two constraints.

(7) *HS+SM representation of a counterbleeding interaction*

Step 1: Palatalization (bleeding candidate blocked by SM constraint)

/esi/	SM(*[si], Agree)	Agree	Ident(High)	*[si]	Ident(Anterior)
[esi]		*		W*	L
ese <BOTH>	W*	L	W*		L
☞ eʃi <*[si]>		*			*

Step 2: Harmony

eʃi <*[si]>	SM(*[si], Agree)	Agree	Ident(High)	*[si]	Ident(Anterior)
[eʃi] <*[si]>		W*	L		
☞ eʃe <*[si], Agree>			*		
esi <*[si]>		W*	L	W*	W*

In counterbleeding, a high-ranked SM constraint rules out the candidate that harmonizes on the first step, since harmonizing satisfies both constraints simultaneously. This blocks the bleeding mapping, causing just palatalization to occur in the first step, then harmony in the second, followed by convergence (not shown).

(8) *HS+SM representation of a counterfeeding interaction*

Step 1: Harmony

/ise/	SM(*[si], Agree)	Agree	Ident(High)	*[si]	Ident(Anterior)
[ise]		W*	L	L	
☞ isi <Agree>			*	*	
iʃe		W*	L	L	W*

Step 2: Convergence (palatalization blocked by SM constraint)

isi <Agree>	SM(*[si], Agree)	Agree	Ident(High)	*[si]	Ident(Anterior)
☞ [isi] <Agree>				*	
iʃi <Agree, *[si]>	W*			L	W*

Counterfeeding can be represented using the same ranking,[1] since the constraint SM(*[si], Agree) blocks palatalization from occuring after harmony in the second step. Thus, in this theory, opaque interactions are captured via high ranking of SM constraints, which require that markedness constraints be satisfied serially in a specified order.

Unlike Stratal OT and HS, the next theory, two-level constraints (McCarthy 1996), represents all interactions in a parallel, one-step mapping. Instead of representing opacity using ordered steps, two-level constraints capture the phenomenon by assigning violations to forms with banned sequences, similar to standard markedness constraints. However, like faithfulness constraints, two-level constraints can reference both the UR and the SR of a form. Here, this will be shown using constraint names like *[A]/B/, which assigns a violation to any form with a sequence of segments on the surface in which the first segment is an [A] and the second segment corresponds underlyingly to a /B/. The tableaux in (9) and (10) demonstrate how this analysis works for the opaque interactions.

(9) *Two-level constraint representation of a counterbleeding interaction*

/esi/	*[s]/i/	Agree	*[si]	Ident(Anterior)	Ident(High)
[esi]	W*	W*	*	L	L
[ese]	W*			L	*
[eʃi]		W*		*	L
☞ [eʃe]				*	*

(10) *Two-level constraint representation of a counterfeeding interaction*

/ise/	*[ʃ]/e/	Agree	*[si]	Ident(Anterior)	Ident(High)
[ise]		W*			L
☞ [isi]			*		*
[iʃe]	W*	W*		W*	L
[iʃi]	W*		L	W*	*

In (9), counterbleeding is captured by ranking the two-level constraint *[s]/i/ higher than Ident(Anterior), which requires palatalization of any surface [s] followed by /i/ underlyingly, regardless of the surface height of the vowel. Similarly, in (10), ranking the two-level constraint *[ʃ]/e/ above *[si] prevents palatalization for segments that were not followed by /i/ underlyingly. Essentially, two-level constraints allow the grammar to condition palatalization on the underlying – rather than surface – height of the vowel.

The final theory treats opaque mappings as exceptional (see, e.g. Sanders 2003), using indexed constraints (Pater 2010) paired with classic OT (Prince & Smolensky 1993). Tableaux demonstrating how such an approach can capture the same counterbleeding and counterfeeding mappings discussed above are shown in (11) and (12), with subscripts designating which constraints are indexed to which morphemes, and indexed

[1] However, note that counterbleeding is again more flexible than counterfeeding, since either of the serial markedness constraints can be ranked high to achieve the former, while a specific serial markedness constraint is required for the latter.

constraints only being violated by forms that include the morphemes they are indexed to. Note that an additional markedness constraint, *[s][-Back], is needed in this analysis to motivate the overapplication of palatalization seen in the counterbleeding case.

(11) *Indexed constraint representation of a counterbleeding interaction*

/es$_0$+i$_1$/	*[s][-Back]$_1$	Agree	*[si]	Ident(Anterior)	Ident(High)
[esi]	W*	W*	*	L	L
[ese]	W*			L	*
[eʃi]		W*		*	L
☞ [eʃe]				*	*

(12) *Indexed constraint representation of a counterfeeding interaction*

/is$_2$+e$_3$/	*Ident(Anterior)$_3$	Agree	*[si]	Ident(Anterior)	Ident(High)
[ise]		W*			L
☞ [isi]			*		*
[iʃe]	W*	W*		W*	L
[iʃi]	W*		L	W*	*

The indexed constraints act exactly like the two-level constraints by favoring the forms that result in opaque mappings over the forms that result in transparent ones. However, the crucial difference between these analyses is that the indexed constraints are conditioned on morpheme identity, rather than any underlying phonological features. This means that in the indexed constraints analysis, opaque mappings are not productive and would not generalize to novel morphemes (although, see Nazarov 2019 for a way of making indexation more productive).

As demonstrated throughout this section, each of these four theories can successfully represent opaque and transparent phonological interactions. This means that they all have the expressive power to capture the four interactions of interest, and all interactions should in principle be learnable in each framework. However, since each makes different assumptions about how the interactions are represented, each presents different learning challenges and potentially makes distinct predictions about what is hard to learn and what should occur in novel words and contexts. The next sections present the modeling work that explores these questions.

4 Simulations

To explore the predictions for each of the theories outlined in §3, we implemented each one as a probabilistic pairwise ranking grammar (Jarosz 2015) and modeled learning using *Expectation Driven Learning* (EDL; Jarosz 2015).[2] Probabilistic pairwise ranking is a way of representing probabilistic constraint-based grammars and can be applied to each of the four frameworks of interest. Likewise, EDL is applicable to all of these frameworks. EDL is straightforwardly extendable to any generative framework with probabilistic parameters (for an application of this learning model in the Principles and Parameters framework, see Nazarov & Jarosz 2017). It was first extended to serial HS by Jarosz (2016), and for the current project, we extended it to a two-level Stratal OT framework. Applying EDL to a new framework simply requires implementing a production module (ie GEN and EVAL) for that framework – EDL learning updates treat the production module as a black box and work exactly the same way across frameworks. While other implementations and learning algorithms have been used in the past for some of the theories of interest (e.g. Staubs & Pater 2016, Nazarov & Pater 2017), we used EDL here since it was consistently able to learn all of the phonological interactions in all of the theories, making it possible to systematically analyze differences in learning rates and generalization to novel forms over the course of learning. All learning models make assumptions about how learning works, but it is important to note that EDL does not in any way 'build-in' the MaxUtil or Transparency biases. The biases that arise during the course of learning are

[2] To download the software used for our simulations, visit https://github.com/gajajarosz/hidden-structure and https://github.com/blprickett/StratalOT_EDL.

consequences of general learning principles and the representations available to the learner, which vary depending on theoretical framework.

4.1 *Training* For all of the simulatons presented in this paper, we ran the models with a learning rate of .05 for 100 passes through the full set of training data. EDL can be run either in batch or online mode, and for these simulations we always used online. For all simulations, initial grammars were completely unbiased, with all constraints tied. Since there is some noisiness in the learning process, we ran each language-theory combination ten times to ensure any observed biases are robust.

The data given to the models used the interactions involving vowel harmony and palatalization presented in §1 and §3. Each language included four kinds of words: those that were faithful to their underlying representation, those that only harmonized, those that only palalatalized, and those that demonstrated one of the four relevant interactions. In (13), examples of each word type are shown.

(13) *Training data examples (grey cells withheld from training)*

	Faithful		**Harmonizing**	**Palatalizing**	**Interacting**	
URs	/aki/	/ase/	/eki/	/asi/	/esi/	/ise/
SRs: Feeding Condition	[aki]	[ase]	[eke]	[aʃi]	X	[iʃi]
SRs: Bleeding Condition	[aki]	[ase]	[eke]	[aʃi]	[ese]	X
SRs: C.F. Condition	[aki]	[ase]	[eke]	[aʃi]	X	[isi]
SRs: C.B. Condition	[aki]	[ase]	[eke]	[aʃi]	[eʃe]	X

As demonstrated by the table above, all four languages have equivalent training data for the faithful, harmonizing, and palatalizing words (i.e. the white cells in the table). However, the interacting words differ depending on the kind of interacting UR that is presented *and* whether the language is opaque or transparent. Only one interacting UR and SR pair type was provided in each language condition – the cells for the pairs not presented in a given condition are indicated in grey with an X. For example, a model acquiring the bleeding language would be trained on interacting forms like /esi/→[ese] that apply the two processes in a transparent ordering to the UR /esi/. This model would never be exposed to the feeding/counterfeeding UR /ise/ *or* the opaque ordering of processes (palatalization, then harmony), shown in the table using white cells. Conversely, a model trained on counterfeeding would see interacting forms like /ise/→[isi], with the processes applying in an opaque ordering to the UR /ise/. The training data for each language included these four types of URs, with variation in the tenseness of vowels, yielding a total of 20 items in each language.[3]

In addition to the training data, the model was given a constraint set at the start of learning that included every constraint that was relevant to the theory being tested. While all theories made use of the four standard constraints—Agree, *[si], Ident(Anterior), and Ident(High)—the rest of their constraint sets differed. When using a stratal analysis, the model was given the four standard constraints, with copies of each appearing in both strata. When testing HS with serial markedness reduction, the four main constraints were joined by the serial markedness constraints SM(Agree, *[si]) and SM(*[si], Agree). The two-level constraints worked similarly, but with *[s]/i/ and *[ʃ]/e/ being used instead. Finally, for the indexed analysis, the model was given the four standard constraints, the markedness constraint *s[-Back], and versions of all five constraints that were indexed to each of the suffix morphemes, for a total of 5 general constraints and 20 indexed ones.[4]

4.2 *Testing* As discussed in §2.1, Prickett (2019) observed both a MaxUtil Bias and a Transparency Bias affecting participants' performance on the same forms across language conditions. A MaxUtil Bias was observed in the acquisition of the palatalization process, since participants' accuracy on forms that only palatalized was higher in the feeding and counterbleeding languages. This is exactly where you might expect to see such a bias, since palatalization is the process that applies at different rates across the various languages. In feeding and counterbleeding languages, palatalization applies in the interacting context as well as in the

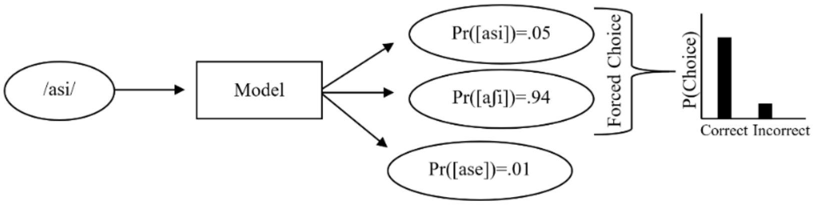
[3] The full set of training data can be downloaded at https://github.com/blprickett/StratalOT_EDL.

[4] While the models were mostly blind to morphological information, for the sake of indexing, some morphology had to be included. For these simulations, each form was broken up into a stem and a suffix, with the latter always being the final segment in the word. For example, the UR /esi/ would be broken up into /es+i/.

palatalizing context, whereas in bleeding and counterfeeding languages, palatalization is only observed in the palatalizing context. Prickett observed the Transparency Bias when comparing accuracy on the interacting forms across language conditions: participants were more accurate in the feeding and bleeding conditions than in the counterfeeding and counterbleeding conditions. The interacting context is where the Transparency Bias arises since that is the context that differs between transparent and opaque language conditions.

To see whether each of the four theories predicts the kind of biases that Prickett (2019) observed in human learning, the models were tested in the same way as his participants. This involved forced choices between either palatalizing or mapping faithfully (in the case of palatalizing forms) and between either a correct or incorrect ordering of the processes (in the case of interacting forms). The forced choice task was simulated by giving the model the appropriate UR (e.g. /asi/) and using the current grammar to estimate the probabilities over all possible output SRs (e.g. [asi], [aʃi], [ase], etc.). Then, the probabilities of the two relevant choices (e.g. [asi] and [aʃi]) were normalized and the normalized probability of the correct SR was treated as the model's accuracy for that choice. This is illustrated in (14). To examine the predictions over the course of learning for each framework, the model's accuracy on each forced choice was calculated after each pass through the training data.

(14) *Illustration of forced-choice task that was given to the model*



To explore the predictions that each theory makes about generalization to novel words, the model was also given a novel UR task. The data for this task consisted of whatever interacting forms the model *wasn't* trained on, given the language it was learning. These novel UR forms are those marked with X in (13), above. For example, when acquiring the bleeding language, the model saw interacting mappings like /esi/→[ese] in its training data. For the novel UR task, it would then be asked to generalize from an interacting UR like /ise/. The model's current grammar was used to estimate the probabilities for each possible SR (as described above for the bias test, but without any normalization) and the SR with the most probability was considered the model's predicted output. As above, the predictions for the generalization task were calculated after each pass through the training data to determine what the model's output was for the majority of acquisition.

5 Results

The results for both the bias test and the novel UR test are described in this section. We considered a bias to be present in any model for which the accuracy across language conditions for the relevant word types exhibited the relative preferences observed in Prickett's study (2019). That is, for MaxUtil Bias, models trained on feeding and counterbleeding had to have higher accuracy on palatalizing items than those trained on bleeding and counterfeeding. Likewise, for Transparency Bias, models trained on bleeding and feeding had to have higher accuracy on interacting items than those trained on the opaque languages. If these relative preferences occurred at any point in learning, the model was considered to have that bias. The results of this analysis are shown in (15).

(15) *Bias Test Results*

Stratal OT	HS+SM	Two-level	Indexed
MaxUtil, Transp.	MaxUtil, Transp.	MaxUtil	MaxUtil

All models were affected by a MaxUtil Bias when performing the forced choice task for palatalizing forms. However, only the serial models (Stratal OT and HS) displayed a Transparency Bias. This means that only the serial theories predict a bias like the one Prickett (2019) observed in his experiment.

The results for the novel UR task were also analyzed and each model's results for a given language were categorized as one of three possible outcomes. The first possibility was that the model predicted a transparent mapping, like /ise/→[iʃi] (i.e., feeding) or /esi/→[ese] (i.e., bleeding). Alternatively, models could predict opaque mappings, like /ise/→[isi] (i.e., counterfeeding) or /esi/→[eʃe] (i.e., counterbleeding). The final option that we observed in the model's output was a faithful mapping, like /ise/→[ise], where no change was made to the novel UR at all. The type of mapping with the highest probability for the majority of the learning,[5] given each theory and language, is shown below in (16).

(16) *Generalization Test Results (underlined labels show which language the model was trained on)*

	Bleeding	Feeding	C.B.	C.F.
Stratal OT	Transparent	Transparent	Transparent	Opaque
HS+SM	Transparent	Transparent	Faithful	Transparent
Two-Level Constraints	Transparent	Opaque	Transparent	Opaque
Indexed Constraints	Transparent	Opaque	Opaque	Opaque

The results above demonstrate that each of the four theories makes unique predictions for the generalization task that we gave to our model. The Stratal model generalizes transparently when trained on bleeding, feeding, and counterbleeding, meaning that it applies feeding, bleeding, and feeding mappings to novel URs in those conditions, respectively. When trained on counterfeeding, the Stratal model generalizes opaquely, meaning that it applies a counterbleeding mapping to the novel UR in that condition. The HS model with serial markedness constraints predicts transparent mappings for novel URs when trained on every language but counterbleeding. For that language, it predicts a faithful mapping, meaning that it would map a novel UR like /ise/ to the SR [ise]. The learners using two-level constraints generalize to novel URs transparently when trained on bleeding or counterbleeding and opaquely when trained on feeding or counterfeeding. And finally, the indexed constraints model predicts opaque generalization (i.e., either counterfeeding or counterbleeding mappings) when trained on every language other than bleeding. When learning bleeding, indexed constraints generalize transparently, meaning that the model predicts a feeding mapping for that novel UR.

6 Discussion

In this paper, we've found novel predictions from four phonological theories that can each represent opaque and transparent interactions. The first set of predictions we showed involved learning biases and demonstrated that all four theories captured the kind of MaxUtil Bias observed by Prickett (2019). This is likely because the maximally utilizing languages (feeding and counterbleeding) provide twice as much evidence for the palatalization process as their counterparts (bleeding and counterfeeding). For example, a model trained on feeding, regardless of which theory it's implementing, will get evidence for ranking *[si] over Ident(Anterior) every time it sees *either* a palatalizing form (e.g., /asi/→[aʃi]) *or* an interacting form (e.g., /ise/→[iʃi]). However, models trained on a bleeding interaction won't get any evidence for the ranking of *[si] and Ident(Anterior) from their interacting forms (e.g., /esi/→[ese]).

While both of the serial models (Stratal and HS) were affected by a transparency bias, neither of the parallel models were. This likely results from the differences in the evidence that each theory needs for a correct ranking of its grammar. Stratal OT and HS both require specific rankings to be learned for the opaque languages that the transparent languages can do without and the model only receives evidence for those rankings when it sees an interacting item in training. For example, to learn counterfeeding, the Stratal OT model must have a ranking in its first stratum that produces just palatalization and a ranking in its second stratum that produces just harmony. Faithful, only-harmonizing, and only-palatalizing items don't give the model any evidence for this, since those processes can all be done in a single stratum's ranking. Only the interacting items (e.g., /ise/→[isi]) show the model that such a specific representation of the pattern is necessary. Similarly, the only evidence for where to rank the serial markedness constraints in HS appears in interacting contexts, and only the opaque languages require a highly specific, high ranking of SM constraints.

The parallel models don't have the same lack of evidence when learning the opaque languages. For example, when acquiring counterbleeding, a two-level constraint model must learn to give *[s]/i/ a high

[5] Note that these were also the results at the end of learning for each model.

ranking. Not only do the interacting forms like /esi/→[eʃe] provide evidence for this ranking, but palatalizing forms like /asi/→[aʃi] support it as well. This means that the model will have relatively high accuracy for interacting items throughout the learning process when it's being trained on the counterbleeding language and explains why no transparency bias appears to affect that model's acquisition. The indexed model works similarly, with the version of *[s][-Back] that's indexed to the underlyingly high suffixes being ranked high due to evidence from both interacting and palatalizing forms.

For the generalization task, none of the theories' predictions are *a priori* obvious. It is only via explicit computational implementation of these theoretical alternatives and analysis of the modeling results that these predictions can be elucidated. The learning pressures are different in each theory because the crucial rankings required to account for the learning data are different in each theory, and the available evidence for the crucial rankings differs across theories and language conditions. We don't have room here to describe why all four theories make the predictions that they do for the novel UR task, but for the remainder of this section, as an example, we'll walk through why the stratal model generalizes in the way that it does. When learning bleeding or feeding, the model *could* represent the full language in a single stratum's rankings, as shown in 0 and (4), since each stratum is essentially its own classic OT grammar. In reality, the model ends up learning these rankings in both of its strata, which maximizes its performance during training, since it gives the model two chances to correctly map each UR to the appropriate SR. When learning counterbleeding, the model needs to learn a ranking that will *only* palatalize in the first stratum, but the second stratum can either apply both changes (palatalization and harmony) or just harmony as shown in (5) and discussed in §3. In our simulations, we found that the model tended to find solutions that did the former, likely because this also maximized performance for palatalizing items in training (by giving the model two chances in the derivation to correctly map those forms). Finally, counterfeeding required the most specific ranking since it needs Stratum 1 to *only* palatalize and Stratum 2 to *only* harmonize, as demonstrated in (6) and explained in §3. The solution that the Stratal model arrives at, given each language, is summarized below in (17) by showing which processes apply in each stratum for each language's grammar.

(17) *Summary of the solutions that the Stratal model arrives at, given training on each language*

	Bleeding	Feeding	Counterbleeding	Counterfeeding
Stratum 1	Pal. & Harm.	Pal. & Harm.	Palatalize	Palatalize
Stratum 2	Pal. & Harm.	Pal. & Harm.	Pal. & Harm.	Harmonize

Each of these solutions affected the model's generalization in different ways (summarized in 18). Since feeding and bleeding led the model to essentially the same grammar, it generalized transparently when given novel URs in both of those cases. For example, when trained on mappings like /esi/→[ese] (i.e. the bleeding language), the model generalized to test URs like /ise/ by putting the most probability on surface forms like [iʃi] (i.e. the feeding candidate). When trained on counterbleeding, the model still generalized transparently to test data like /ise/. This is because the novel UR /ise/ first goes through Stratum 1, which only applies palatalization, and doesn't undergo any changes, since /ise/ doesn't have an /s/ before a high vowel. Then the intermediate form *ise* will pass through the next stratum which applies both processes, resulting in a feeding mapping (as shown in 4), and the SR [iʃi] (i.e., a transparent mapping). Counterfeeding, however, will generalize opaquely, since its test data (e.g., /esi/) will first palatalize (becoming, *eʃi*) and then harmonize (becoming, [eʃe], the counterbleeding candidate).

(18) *Stratal Model's Generalization Summary (changed segments in bold, mapping types italicized)*

	Bleeding	Feeding	Counterbleeding	Counterfeeding
Test UR	/ise/	/esi/	/ise/	/esi/
Stratum 1	i**ʃi**	es**e**	-	e**ʃ**i
Stratum 2	-	-	i**ʃi**	eʃ**e**
Predicted SR	[iʃi] *(Transp.)*	[ese] *(Transp.)*	[iʃi] *(Transp.)*	[eʃe] *(Opaque)*

Future work should explore what humans actually do in a generalization task like the one simulated here. Similar methodologies, in which crucial data is withheld for testing, have been used successfully in a number of past artificial language learning studies (Wilson 2006, Finley 2008, Gallagher 2013 inter alia) to determine what kind of theories best predict phonological generalization in humans. Other avenues for future work could include testing different computational models on this task—Prickett (2019) showed that a neural network could capture MaxUtil and Transparency biases. It could be useful to see whether such a learner's generalization resemble one of the models presented here or if it would also make a unique prediction for the task. Finally, phonological generalization is not a phenomenon that can only be observed in artificial language learning. Some of the most influential studies on generalization in linguistics have used natural language (e.g. Halle 1978) and the novel generalization task we propose here could be applied to speakers of a language with opaque phonology if the correct kinds of interacting forms were absent from that language's lexicon.

While all four of the theories discussed here can capture bleeding, feeding, counterbleeding, and counterfeeding interactions, we found that they all make unique predictions for the tasks we simulated. This work provides a first step toward teasing apart predictions of different theoretical assumptions for learning and generalization. If human behavior can be shown to resemble the predictions made by one of these theories, it could help solve a long-standing debate over which is the appropriate phonological analysis. Furthermore, this work helps demonstrate how implementing phonological theories computationally can help lead to novel predictions and find new ways to determine which theory most closely captures human phonological knowledge and learning.

References

Baković, Eric. (2011). Opacity and ordering. *The Handbook of Phonological Theory, Second Edition* 40–67.

Brooks, K. Michael, Pajak, Bozena, & Baković, Eric. (2013). Learning biases for phonological interactions. *Poster Presented at 2013 Meeting on Phonology*.

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

Finley, Sara. (2008). *Formal and cognitive restrictions on vowel harmony* [PhD Thesis].

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *JML 61*(3) 423–437.

Gallagher, G. (2013). Learning the identity effect as an artificial language. *Phonology 30*(2) 253–295.

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

Hughto, Coral. (2018). Investigating the consequences of iterated learning in phonological typology. *Proceedings of the Society for Computation in Linguistics 1*(1) 182–185.

Jarosz, Gaja. (2015). Expectation driven learning of phonology. *Ms., University of Massachusetts Amherst*.

Jarosz, Gaja. (2016). Learning Opaque and Transparent Interactions in Harmonic Serialism. *Proceedings of AMP 3*.

Jarosz, Gaja. (2014). Serial markedness reduction. *Proceedings of the Annual Meetings on Phonology 1*.

Kim, Yun Jung. (2012). Do learners prefer transparent rule ordering? *Proceedings from CLS 48* 375–386.

Kiparsky, Paul. (1968). Linguistic universals and linguistic change. In Emmon Bach & Robert T. Harms (Eds.) *Universals in linguistic theory* (pp. 170–202). Holt, Rinehart & Winston.

Kiparsky, P. (1971). Historical linguistics. In William Orr Dingwall (Ed.) *A survey of linguistic science* (pp. 576–642).

Kiparsky, Paul. (2000). Opacity and cyclicity. *The Linguistic Review 17*(2–4) 351–366.

McCarthy, John J. (1996). *Remarks on phonological opacity in Optimality Theory*.

McCarthy, John J. (1999). Sympathy and phonological opacity. *Phonology 16*(3) 331–399.

McCarthy, John J. (2000). Harmonic serialism and parallelism. *Linguistics Department Faculty Publication Series* 40.

McCarthy, John J. (2007). *Hidden generalizations: Phonological opacity in Optimality Theory*. Equinox Publishing.

Nazarov, Aleksei. (2019). Formalizing the connection between opaque and exceptionful generalizations. *Toronto Working Papers in Linguistics 41*(1).

Nazarov, Aleksei, & Jarosz, Gaja. (2017). Learning parametric stress without domain-specific mechanisms. *Proceedings of the Annual Meetings on Phonology 4*.

Nazarov, A., & Pater, J. (2017). Learning opacity in Stratal Maximum Entropy Grammar. *Phonology 34*(2) 299–324.

Pater, Joe. (2010). Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker (Ed.) *Phonological argumentation: Essays on evidence and motivation* (pp. 123–154). Equinox Publishing.

Prickett, Brandon. (2019). Learning biases in opaque interactions. *Phonology 36*(4) 627–653. https://doi.org/10.1017/S0952675719000320

Prince, Alan, & Smolensky, Paul. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*.

Sanders, Robert Nathaniel. (2003). *Opacity and sound change in the Polish lexicon* [PhD Thesis].

Stanton, Juliet. (2016). Learnability shapes typology: The case of the midpoint pathology. *Language 92*(4) 753–791.

Staubs, Robert D. (2014). *Computational modeling of learning biases in stress typology*.

Staubs, Robert, & Pater, Joe. (2016). Learning serial constraint-based grammars. In John J. McCarthy & Joe Pater (Eds.) *Harmonic Grammar and Harmonic Serialism* (pp. 369–388). Equinox Publishing.

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