

Learning Opaque and Transparent Interactions in Harmonic Serialism

Gaja Jarosz

University of Massachusetts Amherst

1 Introduction

Based on evidence from linguistic change, primarily cases of diachronic rule reordering, Kiparsky posited two substantive principles governing relative *naturalness* and *learnability* of process interactions. The first hypothesis (Kiparsky 1968) identified *maximal utilization* as the unifying property favoring feeding and counterbleeding orders over bleeding and counterfeeding orders. This proposal met with some criticisms, primarily in the form of evidence suggesting bleeding orders in some cases may be unmarked and in other cases inapplicable as a diagnostic (e.g. Kenstowicz & Kisseberth 1971). In response, Kiparsky (1971) proposed a reformulation, positing instead that *opacity* is what characterizes markedness, unifying counterbleeding and counterfeeding as diachronically disfavored. These hypotheses concern the very nature of phonological systems and have consequences for understanding of typology, learning, and language change. However, exactly how process interactions affect learning and how learning affects language change and typology remain poorly understood. The potential impact on phonological theory of these central issues has been severely limited by a lack of explicit computational models capable of learning opaque interactions and making precise and testable predictions for language acquisition and change.

This paper presents initial modeling results comparing the relative learnability of four basic types of process interactions: bleeding, feeding, counterfeeding and counterbleeding. There are two main contributions. The first is a proof of concept: the paper presents a novel learning model that successfully copes with the hidden structure learning problem instantiated by serial derivations, including opaque derivations, in a variant of Harmonic Serialism. The second contribution is to illuminate some of the learnability consequences of opaque and transparent process interactions, with the principle finding being that ease of learning does not simply depend on the interaction type - how evidence for the processes and their interactions is quantitatively distributed in the learning data is also a principle factor driving learning.

It is not the goal of this paper to argue for a particular approach to opacity. The primary concern of this paper is to analyze and better understand the learning challenges that can arise in these various interactions in as general form as possible. Nonetheless, modeling these interactions and their learning requires that some choices be made about the theoretical framework and learning paradigm, and, as the results will show, these choices have consequences. With all these moving pieces, even apparently simple systems can yield unexpectedly complex learning interdependencies. It is only through the development of explicit computational models capable of learning in the face of hidden structure that we can begin to understand the consequences of these basic theoretical issues for language acquisition, change, and typology.

Section 2 introduces the language system that is used to examine learning of these four interaction types. Section 3 shows how these interactions can be analyzed in the framework of Serial Markedness Reduction (SMR: Jarosz 2014b), a version of Harmonic Serialism (HS) that is capable of representing both transparent and opaque interactions. To model learning in HS/SMR, I extend the Expectation Driven Learning approach of Jarosz (2014a, 2015) to serial derivations. This learning approach is summarized in Section 4. Section 5 then presents the core simulation results, while Section 6 analyzes the underlying factors that give rise to these results. Section 7 concludes.

* I am grateful to Adam Albright, Arto Anttila, Eric Baković, Ryan Bennett, Brett Hyde, Aleksei Nazarov, Joe Pater, Ezer Rasin, and Colin Wilson for helpful comments on this work.

2 Illustrative Example

I focus on a hypothetical system involving potential opaque and transparent interactions between two processes (this example is inspired by a hypothetical example discussed by Baković 2011: 42). One process is *Palatalization* ($s \rightarrow \text{ʃ} / _i$), and the other is *Vowel Deletion* ($V \rightarrow \emptyset / _V$). If deletion occurs first, as illustrated in (1), the interaction is transparent: deletion bleeds palatalization (c), and it also feeds palatalization (d), depending on the configuration in the input. On the other hand, if palatalization occurs first, as shown in (2), the outcome is opaque: deletion counterbleeds palatalization (c) and counterfeeds palatalization (d), depending on the input configuration.

- (1) Transparent Interaction: Deletion both bleeds (1c) and feeds (1d) Palatalization
- | | a. Deletion | b. Palatalization | c. Bleeding | d. Feeding |
|----------------|-------------|-------------------|-------------|------------|
| Underlying | /sua / | si/ | /sia/ | /sui/ |
| Deletion | sa | — | sa | si |
| Palatalization | — | ʃi | — | ʃi |
| Surface | [sa] | [ʃi] | [sa] | [ʃi] |
- (2) Opaque Interaction: Deletion counterbleeds (2c) and counterfeeds (2d) Palatalization
- | | a. Deletion | b. Palatalization | c. Counterbleeding | d. Counterfeeding |
|----------------|-------------|-------------------|--------------------|-------------------|
| Underlying | /sua/ | /si/ | /sia/ | /sui/ |
| Palatalization | — | ʃi | ʃia | — |
| Deletion | sa | — | ʃa | si |
| Surface | [sa] | [ʃi] | [ʃa] | [si] |

This system is obviously schematic, but it involves properties that are characteristic of these interactions in HS in general. By considering a system that can produce both opaque and transparent interactions of the same two processes, it is possible to hold the processes and their analysis constant while varying only their relative ordering in the phonology to examine the consequences of that manipulation. Despite its apparent simplicity, the results will show that this basic system already turns out to involve a number of factors that interact in complex ways during learning.

3 Analysis in Harmonic Serialism with Serial Markedness Reduction



Standard HS with regular markedness and faithfulness constraints cannot model opaque interactions in the general case (see e.g. McCarthy 2000, 2007; Jarosz 2014b), and this holds for the system introduced above as well. As shown in (3)a, palatalization requires that a constraint prohibiting [s] before [i] be ranked above a faithfulness constraint prohibiting changes in place: $*si \gg \text{Ident}$. The tableau in (3)b shows that deletion requires that an anti-hiatus constraint rank above Max: $*VV \gg \text{Max}$.

- (3) a) Palatalization requires $*si \gg \text{Ident}$
- | /si/ | $*si$ | Ident |
|---------|-------|-------|
| a. si | W* | L |
| ☞ b. ʃi | | * |
- b) Deletion requires $*VV \gg \text{Max}$
- | /sua/ | $*VV$ | Max |
|---------|-------|-----|
| a. sua | W* | L |
| ☞ b. sa | | * |


Given the required rankings for the individual processes, neither the counterbleeding nor the counterfeeding derivations can be analyzed in standard HS. Counterbleeding (4) is not possible since the relative rankings necessary for the individual processes predict bleeding on the first iteration. For counterbleeding to occur, palatalization would have to occur on the first pass (candidate c); however, the bleeding candidate (b) satisfies both markedness constraints at the expense of one low-ranked faithfulness constraint. There is no way to rank these constraints to prefer the palatalization candidate since it only satisfies one of the (necessarily high-ranked) markedness constraints, while deletion satisfies both. Counterfeeding is not possible for a different reason. On the first pass, both feeding and counterfeeding require the same output: deletion. As shown in (5), deletion occurs as long as $*VV \gg *si$. In HS, the

candidate with both palatalization and deletion (and therefore neither markedness violation) is not available yet since it requires two operations. This means an intermediate step that creates a violation of **si* is needed. On the second pass, shown in (6), feeding is unstoppable. The **si* \gg Ident ranking necessitates palatalization for an input with a /si/ sequence: with these constraints, there is no way to stop palatalization from occurring once that input configuration is created, as would be required for counterfeeding to occur.



(4) Potential bleeding: Iteration 1

/sia/	<i>*si</i>	Ident	<i>*VV</i>	Max
a. sia	*		*	
 Bleeding	b. sa			*
 Counterbleeding	c. <i>ʃ</i> ia	*	*	

(5) Potential feeding: Iteration 1 requires **VV* \gg **si*


/sui/	<i>*si</i>	Ident	<i>*VV</i>	Max
a. sui	L		W*	L
 Deletion	b. si	*		*

(6) Feeding only: Iteration 2


/si/	<i>*si</i>	Ident	<i>*VV</i>	Max
 Counterfeeding	a. si	*		
 Feeding	b. <i>ʃ</i> i	*		

Serial Markedness Reduction (SMR; Jarosz 2014b) provides a way to analyze these opaque interactions. SMR relies on the intuition that processes in OT and HS correspond to satisfaction of markedness constraints, and it utilizes two extensions to the standard HS framework to track and evaluate the order in which markedness constraints are satisfied in a derivation. The first extension is a list called *Mseq* that tracks satisfaction of markedness constraints during the construction of the derivation. *Mseq* is part of the candidate, it gets updated each iteration based on which constraints are satisfied as compared to the faithful candidate, and then it gets passed down to the next iteration as part of the input. This is shown in (7)a for palatalization: the *Mseq* starts out empty $\langle \rangle$ in the input and acquires **si* as soon as palatalization occurs. In (7)b, the same occurs for deletion: **VV* is added to the *Mseq* for the candidate that satisfies it. On the next pass, the *Mseqs* are provided as part of the input and are updated for candidates that eliminate additional markedness violations by appending those markedness constraints to the end of the list.

(7) a) Palatalization satisfies **si*

/si/ $\langle \rangle$	<i>*si</i>	Ident
a. si $\langle \rangle$	W*	L
 b. <i>ʃ</i> i $\langle *si \rangle$		*

b) Deletion satisfies **VV*

/sua/ $\langle \rangle$	<i>*VV</i>	Max
a. sua $\langle \rangle$	W*	L
 b. sa $\langle *VV \rangle$		*

The second extension of SMR is a novel type of constraint (a Serial Markedness, or SM, constraint) that examines the order in which markedness constraints are satisfied in a candidate's *Mseq*. SM(M_1 , M_2) favors derivations where M_1 is satisfied before M_2 : it penalizes candidates whose *Mseqs* have M_2 preceding or occurring simultaneously with M_1 . These are the essential building blocks of SMR; for further technical details, including some proposed restrictions on the application of SM constraints, see Jarosz (2014b).

The presence of the SM constraints does not affect the analysis for non-interacting inputs /si/ and /sua/ since no candidates in these derivations satisfy both markedness violations. These extensions do make it possible, however, to analyze both bleeding and counterbleeding, with the relative ranking of SM(**si*, **VV*) determining the outcome, as shown in (8). SM(**si*, **VV*) assigns a violation to the bleeding candidate (b) because it simultaneously satisfies **si* and **VV* rather than in the specified order. If SM(**si*, **VV*) ranks above **VV* and Ident, counterbleeding occurs; otherwise, bleeding occurs.

Both feeding and counterfeeding are now possible as well, depending on the ranking of the SM constraint. Deletion applies on the first pass through the grammar as before (9), and the presence of **VV* in

the winner's *Mseq* reflects this. On the second pass through the grammar (10), the *Mseq* of the input now indicates this input has already undergone deletion, and $SM(*si, *VV)$ therefore penalizes the transparent feeding candidate (b) whose *Mseq* encodes the fact that palatalization follows deletion. This makes it possible to capture counterfeeding if $SM(*si, *VV) \gg *si$ holds and feeding if the opposite ranking holds.

- (8) Potential bleeding: Iteration 1 counterbleeding iff $SM(*si, *VV) \gg \{Ident, *VV\}$

/sia/ <>	<i>*si</i>	Ident	<i>*VV</i>	Max	$SM(*si, *VV)$
a. sia <>	*		*		
B b. sa < <i>*si</i> + <i>*VV</i> >				*	*
CB c. <i>ʃ</i> ia < <i>*si</i> >		*	*		

- (9) Potential feeding: Iteration 1 requires $*VV \gg *si$

/sui/ <>	<i>*si</i>	Ident	<i>*VV</i>	Max	$SM(*si, *VV)$
a. sui <>	L		W*	L	
b. si < <i>*VV</i> >	*			*	

- (10) Counterfeeding: Iteration 2 requires $SM(*si, *VV) \gg *si$

/si/ < <i>*VV</i> >	<i>*si</i>	Ident	<i>*VV</i>	Max	$SM(*si, *VV)$
CF a. si < <i>*VV</i> >	*				
F b. <i>ʃ</i> i < <i>*VV, *si</i> >		*			*

4 The Learning Problem & Approach

Learning a constraint ranking in HS presents a hidden-structure learning problem since the learner is not given information about all input-output mappings. The learner may know that underlying form /sia/ maps to surface [ʃi] but does not know what intermediate representations (e.g. winners), if any, may be required to generate this mapping (see Tessier & Jesney 2014 and Merchant 2014 for related discussion). A great deal of recent work addresses learning of various kinds of hidden structure, including the ambiguity created by abstract prosodic structure and underlying representations (see Jarosz in press for an overview).

To learn HS derivations, I extend the Expectation Driven Learning (EDL) approach that Jarosz (2014a, 2015) applies to learning of hidden metrical structure and underlying representations¹. EDL is a probabilistic learner and relies on a noisy ranking represented in terms of pairwise ranking probabilities. These noisy rankings are a stochastic generalization of Partially Ordered Grammars (Anttila 1997).

- (11) Stochastic Pairwise Ranking Grammar

	A	B	C	D
A		1	1	0
B	0		.7	0
C	0	.3		0
D	1	1	1	

The grammar can be represented using a table such as the one in (11), where each cell shows the chance of ranking one constraint above another. Just half of the table is necessary to specify the grammar since corresponding values across the diagonal must sum to 1 and are redundant. The pairwise ranking probabilities are used when sampling a total ranking for production, which is done by iteratively setting cells in the table to 0 or 1 until a total ranking is specified (see Jarosz 2015 for details).

Learning in EDL is formalized as a probabilistic inference problem about these pairwise ranking probabilities. On each update, the task of the learner is to determine for each pair of constraints A, B, the

¹ Due to space limitations, the description of the learning model presented here is just an overview of the basic approach. See Jarosz (2015) for technical details. For other recent approaches to learning serial derivations see Tessier & Jesney (2014), Staubs & Pater (2016), Nazarov & Pater (2013), and Merchant (2014).

probability that A ranks above B, given the observed data form d and the current grammar g : $\Pr(A \gg B|d, g)$. That is, the learner’s task is to make inferences about the pairwise ranking given the data. As shown in (12), it is possible to flip the conditioning in this expression using Bayes’ Law into quantities associated with *production of overt data* d given the current grammar g and the ranking of $A \gg B$.

(12) Pairwise Ranking Inference

$$\Pr(A \gg B|d, g) = \frac{\Pr(d|A \gg B, g) \Pr(A \gg B|g)}{\Pr(d|A \gg B, g) \Pr(A \gg B|g) + \Pr(d|B \gg A, g) \Pr(B \gg A|g)}$$

The learner exploits this decomposition by using its production module to test each pairwise ranking’s consequences for production. Jarosz shows can be done by constrained sampling from the production module, temporarily setting each pairwise ranking one-by-one (while leaving the rest of the grammar unchanged) and testing how often that pairwise ranking generates a matching output. Intuitively, the pairwise ranking that more often generates the right surface output compared to its reverse is more successful. This sampling procedure provides an estimate of $\Pr(d|A \gg B, g)$, which is then plugged into (12). The other quantity in the numerator is simply the probability of $A \gg B$ in the current grammar, and the denominator is just the sum of these quantities for both relative rankings. These are used to estimate $\Pr(A \gg B|d, g)$ for each overt data form, and the batch learning update used in these simulations computes this for each data form and combines the results in a single update for each pair of constraints.

The main take-home points about EDL’s learning strategy is that the learner tests each pairwise ranking one-by-one on the data and increases the probability of pairwise rankings that are better at generating the data. The better a relative ranking is than its reverse at generating the overall patterns in the data, the more that pairwise ranking probability is rewarded. Learning is sensitive to frequency: the more data a relative ranking successfully explains, the larger the update. The learner never explicitly grapples with hidden structure, instead relying only on its production module’s successes and failures in matching the overt data when various pairwise rankings are tested. Consequently, the learner can effectively treat the production module as a black box. Because of this generality, extending this learning approach to HS derivations is straightforward; it requires only that an HS EVAL module be used for production.

5 Simulations

The simulations make the fewest possible assumptions and incorporate no explicit biases favoring any of the process interactions. All simulations assume an unbiased initial ranking, with all constraints tied (this corresponds to a table with all pairwise probabilities set to 50%). For symmetry, both SM constraints are included in the full set of constraints so that there are constraints preferring each ordering of the processes available and competing in the constraint set. The full constraint set therefore is **s_i*, **VV*, *Max*, *Ident*, *SM(*s_i*, **VV*), and *SM(*VV*, **s_i*). This means there is no bias for either the transparent or opaque order inherent to constraint set itself.

The first set of simulations examine whether EDL is capable of learning both transparent and opaque interactions. This is important since any conclusions about relative ease of learning are only meaningful if the learning framework is capable of learning both in principle. To test EDL’s ability to learn these complete systems, the learner was presented with two sets of learning data that provide complete information about all four underlying to surface form mappings, as shown in (13). The learner still has to grapple with hidden intermediate representations but is provided with explicit information about the existence of the individual processes of palatalization and deletion as well as the UR to SR mappings for the two input configurations that result in interactions, /sia/ and /sui/. Learning is nondeterministic so the simulations were repeated twenty times for each language to examine the reliability of the results.

(13) Learning Data for Complete Systems

a) Transparent system	b) Opaque system
/si/ → [ʃi]	/si/ → [ʃi]
/sua/ → [sa]	/sua/ → [sa]
/sia/ → [sa]	/sia/ → [ʃa]
/sui/ → [ʃi]	/sui/ → [si]

The model succeeds at learning both systems on all runs, which shows that it can handle learning in HS/SMR in principle. On average it takes longer to converge on the grammar for the opaque system than for the transparent system (24.1 vs. 20.5 iterations; $p < 0.01$). While this suggests there may be something harder to learn about the opaque system as a whole than the transparent system as a whole, this does not make it possible to compare the learning of the four interaction types (since each system involves two).

To determine whether any learning biases emerge that are specific to particular process interactions, the second series of simulations examine what happens when the learner is only asked to learn one process interaction at a time. In this case, the learning data for each “single interaction” system, summarized in (14), provides the learner with explicit information that palatalization and deletion occur. Additionally, each set of learning data provides explicit evidence of one other UR to SR mapping, corresponding to one of the interacting contexts. For example, in the bleeding single interaction system (a), the learner has information that both palatalization and deletion occur and that there is a bleeding interaction for input /sia/. The learner is not constrained by any other learning data, which makes it possible to determine whether some of the process interactions are inherently more difficult to learn than others when the learner has unambiguous evidence of the individual processes.

(14) Learning Data for Single Interaction Simulations

a) Bleeding	b) Feeding	c) Counterbleeding	d) Counterfeeding
/si/ → [ʃi]	/si/ → [ʃi]	/si/ → [ʃi]	/si/ → [ʃi]
/sua/ → [sa]	/sua/ → [sa]	/sua/ → [sa]	/sua/ → [sa]
<u>/sia/ → [sa]</u>	<u>/sui/ → [ʃi]</u>	<u>/sia/ → [ʃa]</u>	<u>/sui/ → [si]</u>

As discussed above, learning updates are influenced by how often the patterns supporting particular relative rankings are instantiated in the learning data. To investigate the role input frequency may play in learning process interactions, I considered three variants of each single interaction system, differing in how often the individual processes occur relative to the interacting context (15). In the “low” interaction distribution (a), the individual processes occur ten times more frequently than the interacting context. In the “uniform” interaction distribution (b), the individual processes and the interacting contexts occur equally often. Finally, in the interaction “high” distribution (c), the interacting context occurs ten times more often than the individual processes.

(15) Learning Data Input Distributions for Single Interaction Simulations

a) Interaction “Low”	b) Interaction “Uniform”	c) Interaction “High”
Palatalization: 10	Palatalization: 10	Palatalization: 1
Deletion: 10	Deletion: 10	Deletion: 1
Interaction: 1	Interaction: 10	Interaction: 10

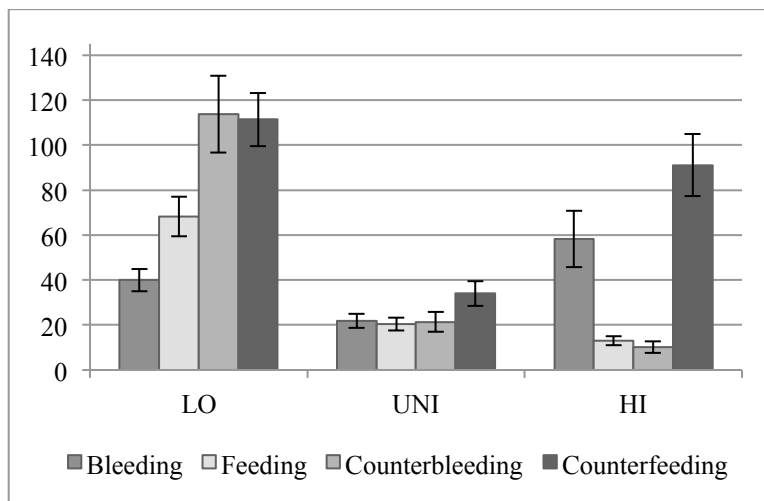


Figure 1 – Number of Iterations till Convergence (error bars show standard deviation)

Since there are four systems with three distributions each, there are twelve sets of learning data. As before, the learning simulations were repeated 20 times for each combination to ensure any observed learning differences were reliable. The average number of iterations the learner required till convergence (and the standard deviation) is shown for all twelve systems in Figure 1. As is evident from the figure, the relative learning rates for the four process interactions depend on the input distribution. There are three main generalizations about the results, which the next section examines in turn. First, in the low interaction distribution, the learner takes longer to learn the opaque process interactions than the transparent interactions, consistent with Kiparsky's (1971) second hypothesis identifying transparency as the crucial dimension. Second, in the high interaction distribution, counterfeeding and bleeding are learned most slowly, consistent with Kiparsky's (1968) earlier hypothesis that acquisition favors maximal utilization of processes. Finally, feeding is learned more slowly than bleeding in the low interaction distribution.

6 Analysis & Discussion

There are no built-in biases in this learning model, constraint set, or starting grammars – so where are these biases coming from? This section analyzes the learners' behavior in more detail to clarify this rather complex pattern of results. There are three interacting factors affecting learning in these simulations, one corresponding to each main observation above, and each is addressed in turn in the following subsections.

6.1 Amount of Learning Required The first asymmetry concerns the preference for transparent over opaque process interactions in the low distribution context. This asymmetry derives from the nature of unfaithful mappings in HS, the required rankings for SM constraints in each kind of interaction, and the gradualness of constraint reranking in this learning model. The analysis in Section 3 showed that SM constraints have to be higher ranked than the markedness constraints to get opaque interactions, while the individual processes require that the markedness constraints rank higher than the faithfulness constraints. This means that the opaque interactions require three fully separated 'strata' of rankings, with SM constraints at the top, some markedness constraints crucially below them, and faithfulness constraints below markedness constraints. In contrast, the transparent interactions require that the markedness constraints be higher ranked than the faithfulness and SM constraints, but no particular ranking is necessary between the SM and faithfulness constraints.

Since the individual processes are frequent in the low interaction learning data, the learner is given reliable and frequent evidence that each markedness constraint must be ranked above its corresponding faithfulness constraint. All learners in the low interaction context therefore quickly learn that both palatalization and deletion occur. To get the opaque interactions right, learners additionally have to rank the SM constraints fully above the markedness constraints. The crucial evidence for ranking SM constraints above markedness constraints, however, is very rare since SM constraints are only relevant for generating the output in the interacting context, which occurs infrequently. As a result, it takes a long time for the learner to establish a categorical ranking with the SM constraints strictly above the markedness constraints. The learner passes through an intermediate stage where deletion and palatalization are essentially categorically applied in the appropriate, non-interacting contexts, but the processes are ordered variably in the interacting context. This corresponds nicely with Kiparsky's discussion of diachronic rule reordering: the simulations indicate that the difficulty in learning these interactions lies in difficulty learning the ordering itself. The transparent interactions take less time to learn because the SM constraints do not need to be placed in their own separate stratum above the markedness constraints. Essentially, the SM constraints do not have as far to travel to reach their target ranking capturing the transparent interactions.

Overall, therefore, the preference for transparent over opaque interactions in the low distribution context is in part due to a difference in the total amount of learning that is required of the learner in each case, in particular, the distance that the SM constraints have to travel during learning. The low frequency of the data providing information about the ranking of SM constraints is also involved. The three-strata opaque grammar is more distant from the initial tied constraint ranking than is the transparent grammar that can rank SM constraints together with faithfulness. Note that this notion of 'amount of learning' is distinct from several related alternatives. It is not directly related to the number of total orders of constraints corresponding to each target language: bleeding (80) > feeding (48) > counterfeeding (42) > counterbleeding (28). Picking total orders at random from a tied grammar would predict that

counterbleeding should be disadvantaged compared to counterfeeding, and this is not the observed behavior of the model. This pattern of results is also not directly attributable to how easily the initial tied grammar can generate the correct output in the interacting context: bleeding (30%) > counterfeeding (27%) > feeding (7%) > counterbleeding (4%). Instead, the time till convergence depends on how quickly the learner can establish a reliable ranking for the SM constraints, for which the evidence is rare.

6.2 Quality of Evidence When the interacting contexts are frequent, on the other hand, learning is more sensitive to the crucial ranking information the outputs in the interacting contexts reveal. Closer inspection reveals that the counterbleeding and feeding interaction contexts unambiguously require all the crucial rankings for the target language, while the bleeding and counterfeeding contexts obscure crucial rankings about the individual processes or their interaction or both. When the learner is confronted with reliable evidence about all the rankings frequently, they learn quickly (feeding, counterbleeding), but when the majority of the learning data provides ambiguous or incomplete information about the necessary rankings (bleeding, counterfeeding), learning occurs more slowly.

In counterbleeding and feeding interactions, both processes successfully apply, providing the learner with reliable evidence that the markedness constraints must be ranked above the faithfulness constraints. The only way the learner can generate the correct outcomes in these cases is to posit both individual processes. Furthermore, the learner also has to posit the correct ranking of the SM constraints to make sure that opaque outcome is favored for counterbleeding and the transparent outcome is favored for feeding.

This is shown in (16) and (17) for counterbleeding. To get palatalization on the first step, **si* has to rank above Ident, and one of the SM constraints has to block the bleeding candidate. When the learner tests the relative ranking of **si* vs. Ident, only **si* >> Ident can produce the correct output: there is no way to generate the observed output with the reverse ranking. Since the learner gets this evidence often, the learner very quickly begins to favor **si* >> Ident in the grammar. Similarly, to get deletion on the second iteration, **VV* has to rank above Max and SM(**VV,*si*). When the learner tests the reverse rankings, they cannot generate the correct output and are penalized accordingly. Every time the learner is presented with the mapping /sia/ → [ʃa], they are given reliable information about the rankings required for the language as a whole. Because this information is provided frequently, the learner quickly acquires the target language.

(16) Counterbleeding: Iteration 1 – Evidence for Palatalization and SM ranking

/sia/ <>	<i>*si</i>	Ident	<i>*VV</i>	Max	SM(<i>*si,*VV</i>)	SM(<i>*VV,*si</i>)
a. sia <>	W*	L	*			
b. sa < <i>*si+*VV</i> >		L	L	W*	W*	W*
☞ c. ʃia < <i>*si</i> >		*	*			

(17) Counterbleeding: Iteration 2 – Evidence for Deletion and SM ranking

/ʃia/ < <i>*si</i> >	<i>*si</i>	Ident	<i>*VV</i>	Max	SM(<i>*si,*VV</i>)	SM(<i>*VV,*si</i>)
a. ʃia < <i>*si</i> >			W*	L		L
☞ b. ʃa < <i>*si,*VV</i> >				*		*

In the feeding interaction, shown in (18) and (19), the situation is similar. To get the right output, the learner must posit the rankings to get deletion on the first iteration and palatalization on the second iteration, which includes correctly ranking the SM constraint so that counterfeeding does not occur. Rankings that are incorrect for the target language, such as **si* >> **VV*, have no chance of generating the observed output for the input /sui/, and every time the learner is presented with this input, the correct rankings are reinforced.

(18) Feeding: Iteration 1 – Evidence for Deletion

/sui/ <>	<i>*si</i>	Ident	<i>*VV</i>	Max	SM(<i>*si,*VV</i>)	SM(<i>*VV,*si</i>)
a. sui <>	L		W*	L		
☞ b. si < <i>*VV</i> >	*			*		

(19) Feeding: Iteration 2 – Evidence for Palatalization and SM ranking

/si/ <*VV>	*si	Ident	*VV	Max	SM(*si,*VV)	SM(*VV,*si)
a. si <*VV>	W*	L			L	
b. ʃi <*VV,*si>		*			*	

The ranking evidence available in the interacting bleeding and counterfeeding contexts looks quite different. Bleeding and counterfeeding both obscure the evidence of palatalization. These input-output mappings are equally compatible with a grammar that lacks palatalization altogether. They provide evidence of deletion, but are ambiguous with respect to the existence of palatalization (and therefore the ranking of *si) and the potential interaction of palatalization and deletion (and therefore the ranking of the SM constraints). Indeed, in the bleeding case (20), even the target relative ranking of *VV \gg Max is ambiguous: it is possible to achieve the observed output under either relative ranking. Thus, the bleeding interacting context alone provides little information to the learner; it is only by examining the outcomes for other inputs that the learner can ultimately determine the correct target grammar.

(20) Bleeding: Iteration 1 – Only Ambiguous Evidence for Deletion

/sia/ <>	*si	Ident	*VV	Max	SM(*si,*VV)	SM(*VV,*si)
a. sia <>	W*		W*	L	L	L
b. sa <*si+*VV>				*	*	*
c. ʃia <*si>		W*	W*	L	L	L

For counterfeeding (21)-(22), deletion requires that *VV rank above Max and *si - otherwise the predicted outcome is the faithful candidate. However, ranking of either Ident or SM(*si,*VV) above *si can explain why the derivation converges on [si]. In other words, lack of palatalization predicts the same result, and it is only by consulting the rarely-occurring inputs demonstrating the individual processes that the learner can disambiguate these possibilities.

(21) Counterfeeding: Iteration 1 – Evidence for Deletion

/sui/ <>	*si	Ident	*VV	Max	SM(*si,*VV)	SM(*VV,*si)
a. sui <>	L		W*	L		
b. si <*VV>	*			*		

(22) Counterfeeding: Iteration 2 – Ambiguous Evidence for SM ranking

/si/ <*VV>	*si	Ident	*VV	Max	SM(*si,*VV)	SM(*VV,*si)
a. si <*VV>	*					
b. ʃi <*VV,*si>	L	W*			W*	

As discussed in the preceding section, in the “low” interaction distribution, the learner acquires deletion and palatalization quickly and takes longer to settle on the correct ranking of the SM constraints. The ordering is what takes time to learn. In contrast, in the “high” interaction distribution, learning the interaction is not the bottleneck: learning palatalization is what takes a long time, and learning the ranking of the SM constraints proceeds mostly in parallel as the learner slowly determines that *si \gg Ident. In this context, therefore, learning difficulty is more suggestive of rule loss than re-ordering.

The comparison between these two cases reveals a trade-off: highly-constraining learning data can be very informative for the learner when it occurs frequently, but when it is the sole and infrequent evidence for a ranking requirement in the target language, its effects can take time to accrue. This is sensible behavior for a statistical learner who must treat all learning data with some caution and skepticism: infrequent patterns, which could be due to noise, errors, or variation, should not cause the learner to dramatically restructure their grammar. Ranking evidence that is more frequently available in the learning data is learned more quickly, and in the present learning context, maximal utilization provides completely reliable evidence of the target ranking.

6.3 Other Constraint Interactions Finally, the results also show that the model learns bleeding interactions more quickly than feeding interactions in the “low” interaction distribution. This does not follow from either of Kiparsky’s hypotheses. Indeed, Kiparsky claimed that ‘the unmarked status of feeding order is not subject to any serious doubt’ (1971: 612). It is important to understand where this prediction comes from since the predictions are otherwise reasonably expected and compatible with Kiparsky’s hypotheses.

Recall that in this distribution, the learner has frequent and unambiguous evidence of both palatalization and deletion. The SM constraints can be low-ranked for both feeding and bleeding interactions. Why, then, would the learner require more time to settle on a feeding interaction? It turns out that this prediction follows from a rather surprising and likely undesirable typological prediction of HS. Feeding and counterfeeding both require deletion on the first iteration. As repeated in (23), for deletion to be optimal, a ranking is required between the two markedness constraints: satisfaction of *VV creates a violation of *si so deletion requires *VV >> *si. Under the opposite relative ranking, a pattern I will refer to as pre-emptive blocking is predicted: palatalization occurs for /si/ → [ʃi], deletion occurs for /sua/ → [sa], bleeding occurs for /sia/ → [sa], but for the input /sui/, where feeding would be expected, the mapping is faithful: deletion is pre-emptively blocked from occurring in just those cases where it would feed palatalization if it did occur. This is a prediction of standard HS, not SMR, but SMR inherits it.

(23) Feeding and Counterfeeding require *VV >> *si on Iteration 1

/sui/	*si	Ident	*VV	Max
a. sui	L		W*	L
☞ b. si	*			*

A complete solution to this typological question is beyond the scope of this paper. My primary concern at present is to show that this prediction does indeed follow from this property of potential feeding interactions in HS. This prediction means that feeding requires an additional relative ranking be learned that bleeding does not, and it is learning this additional ranking between the two markedness constraints that slows down the learning of feeding as compared to bleeding interactions. To verify this I show that if the definition of the constraints are modified so as to eliminate this typological prediction (and this required ranking), the learning advantage of bleeding over feeding disappears. It is possible to rule out this typological prediction and this crucial ranking by modifying the definition of *si. Previous work in HS identifies advantages to defining the conditioning context or position in the input (Jesney 2011). If *si is replaced by a two-level constraint *s/_i that penalizes a surface [s] whose input correspondent precedes an input /i/, violations of this constraint do not show up on the iteration that the [si] sequence is created, shown in (24a)². The violation, shown in parentheses, is not there (yet) because the surface [s]’s input correspondent is not followed by /i/ in the input. Because the violation is not really there, there is no ranking requirement between *VV and *s/_i. On the next iteration, shown in (24b), the input has a /si/ sequence, however, and so the surface [s] now violates this constraint and palatalization can occur.

24) a) No blocking, only feeding: Iteration 1

/sui/	*s/_i	Ident	*VV	Max
a. sui			W*	
☞ b. si	(*)			L*

b) No blocking, only feeding: Iteration 2

/si/	*s/_i	Ident	*VV	Max
a. si	W*			
☞ b. ʃi		L*		

My goal here is not to argue for this approach to typology in HS, but rather to show that when the typological predictions are altered, this affects the learning predictions. This is demonstrated in **Figure 2**, which repeats the single interaction simulations using the new constraint instead of *si. The results reveal qualitatively identical patterns, except that feeding and bleeding are now learned at comparable rates in the low interaction distribution. I therefore conclude that it is the ranking requirement between markedness

² This approach has similarities to Comparative Markedness (McCarthy 2003) and Previous Step (Kavitskaya & Staroverov 2010; Staroverov 2010) constraints.

constraints in HS in potentially feeding interactions that is responsible for the learning advantage observed for bleeding over feeding in the earlier simulations.

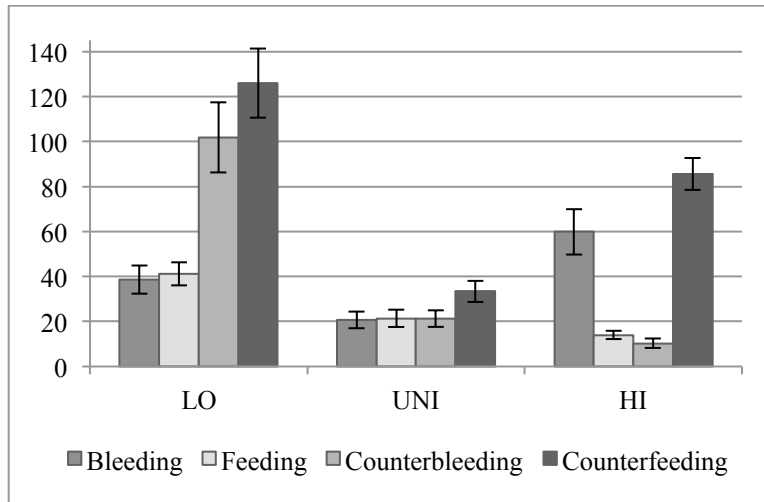


Figure 2 – Number of Iterations till Convergence with **s/_i*

7 General Discussion

It is not straightforward to connect diachronic tendencies with learning difficulty, even with an explicit model of learning difficulty and a simple, artificial language system. The simulations presented here do not predict an absolute hierarchy of learning difficulty for the four interaction types. Rather, relative learning difficulty depends on numerous factors, including how evidence for the target language is distributed in the learning data, how distant the target language is from the initial grammar, and what alternative hypotheses the learner must rule out during learning. While some difficulties in learning are due to general characteristics of the learning data (frequency) and general properties of constraint interaction in HS (high or low ranking of SM constraints), it is also possible for relatively minor changes in the definition of a single constraint to affect the overall predictions as well. This underscores the importance of developing and thoroughly exploring the predictions of computational learning models that implement various theoretical assumptions about how phonological systems work. Generating predictions for learning and ultimately for language change requires fully explicit and testable models of learning and of phonology.

Investigating these questions with more realistic linguistic systems is an obvious and essential next step, but analyzing the behavior of the model in manageable systems like this one is crucial to identifying and understanding the general properties of these interactions and the learning data that may affect learning. The analyses presented here take a first step toward identifying some of the crucial factors that may be involved. Perhaps the most significant and general of these factors is the question of how evidence for the target ranking is distributed in the learning data. These simulations make the simplifying assumption that underlying representations are available to the learner. This is not necessary for successful learning in this framework; however, it does conveniently make it possible to independently manipulate how frequent evidence for the individual processes and their interaction is in the input and to examine the consequences of these manipulations. In a more realistic scenario, no single form can provide unambiguous information about the target grammar since the evidence for the processes and their interaction is only evidenced by alternations across multiple forms. This makes the question of how reliable and readily available evidence for the target grammar is under various conditions in real languages all the more complex (and interesting).

If the predictions of this model are on the right track and learning does in fact depend on how often various patterns are instantiated in the learning data, this means that predictions for language change will crucially depend not only on the inherent nature of the interactions in the target grammar but also on how these interactions are evidenced quantitatively in the available data. The quantitative patterns themselves depend not only on the phonological processes involved but also on the morphological processes that give rise to the various contexts of application. This highlights a direction for further research on the diachronic

stability and instability of various process interactions: to what extent can seemingly divergent diachronic outcomes be understood once the reliability of the evidence for particular processes and their interactions in these languages is taken into account? For example, does the stability of the famous opaque interaction between raising and flapping in Canadian English depend on the productivity of the suffixes –er, –ing and others that reliably provide learners with evidence about the interaction of the two processes?

Another significant prediction of this model is that difficulty learning a process interaction may be due to problems learning the interaction *per se* or to learning one of the processes involved, with different predictions for diachronic change expected in each case. These distinctions line up with Kiparsky's (1971) discussion of diachronic rule re-ordering and rule loss; however, the predictions of this model are that re-ordering and loss may arise under different circumstances and for different reasons. The difficulty of learning opaque interactions in the low interaction distribution is caused by difficulty learning the appropriate ordering between easily acquired individual processes, while the learning difficulty in the high interaction distribution stems from difficulty learning one of the processes itself. If these simulations scale up to more realistic and complex learning systems, it would be reasonable to expect that incomplete learning in the former case may gradually lead to re-ordering, while incomplete learning in the latter case may lead to loss.

In sum, analysis of the simplest possible system capable of modeling each of these interactions already paints a complex picture with multiple factors interacting during learning and multiple ways for learning to be difficult. Further understanding of these and other constraints on learning and predictions for diachronic change will require closer examination of how both human and machine learning of interacting processes depends on the quantitative properties of the language input.

References

- Anttila, Arto. 1997. Deriving variation from grammar. In Frans Hinskens, Roland van Hout and Leo Wetzels (eds.), *Variation, Change and Phonological Theory*. Amsterdam, John Benjamins. 35-68.
- Baković, Eric. 2011. Opacity and ordering. *The Handbook of Phonological Theory* (2nd ed). Ed. John Goldsmith, Jason Riggle, and Alan Yu. Malden, MA: Wiley-Blackwell. 40-67.
- Jarosz, Gaja. 2015. Expectation Driven Learning of Phonology. Ms., University of Massachusetts Amherst.
- Jarosz, Gaja. 2014a. Stochastic, reward-based learning of hidden structure in phonology. Paper presented at 11th Meeting of the Old World Conference in Phonology, Leiden, Holland.
- Jarosz, Gaja. 2014b. Serial Markedness Reduction. *Proceedings of the 2013 Meeting on Phonology 1(1)*, Amherst, MA.
- Jarosz, Gaja. In press. Learning with Violable Constraints. In Jeff Lidz, William Snyder, & Joe Pater (eds.), *The Oxford Handbook of Developmental Linguistics*. Oxford University Press.
- Jesney, Karen. 2011. Positional Faithfulness, non-locality, and the Harmonic Serialism solution. In Suzi Lima, Kevin Mullin & Brian Smith (eds.), *Proceedings of the 39th Meeting of the North East Linguistics Society (NELS 39)*, 403-416. Amherst, MA: GLSA.
- Kavitskaya, Darya, and Staroverov, Peter. 2010. When an interaction is both opaque and transparent: the paradox of fed counterfeeding. *Phonology* 27(02). Cambridge: Cambridge University Press. 255-288.
- Kenstowicz, Michael, and Charles Kisseberth. 1971. Unmarked bleeding orders. *Studies in the Linguistic Sciences* 1(1): 1-12.
- Kiparsky, Paul. 1968. Linguistic universals and linguistic change. In Emmon Bach & Robert T. Harms (eds.) *Universals in linguistic theory*. New York : Holt, Reinhart & Winston. 170–202.
- Kiparsky, Paul. 1971. Historical linguistics. In W. O. Dingwall (ed.) *A Survey of Linguistic Science*. College Park: University of Maryland Linguistics Program. 576-642.
- McCarthy, John J. 2000. *Harmonic serialism and parallelism*. GLSA.
- McCarthy, John J. 2003. Comparative markedness. *Theoretical Linguistics* 29: 1-51.
- McCarthy, John J. 2007. *Hidden generalizations: phonological opacity in Optimality Theory*. Equinox.
- Merchant, Nazarré. 2014/Under review. Recursive Join Learning with Candidate Maps.
- Nazarov, Aleksei, and Pater, Joe. 2013. Learning opacity in a stratal Maximal Entropy framework. Poster presented at the 21st Manchester Phonology Meeting, University of Manchester.
- Staroverov, Peter. 2010. Too-many-solutions and Reference to Position in Serial OT. *University of Pennsylvania Working Papers in Linguistics*: Vol. 16(1), Article 23.
- Staub, Robert, and Pater, Joe. To appear 2016. Learning serial constraint-based grammars. In John McCarthy and Joe Pater (eds.) *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.
- Tessier, Anne-Michelle, and Jesney, Karen. 2014. Learning in Harmonic Serialism and the necessity of a richer base. *Phonology* 31(1): 155-178.