Learning biases in opaque interactions

1. Introduction

Artificial language learning studies have been used extensively to explore cognitive biases in phonological acquisition, with the goal of determining which biases a theory of phonology should predict. Two categories of bias have dominated this line of research: substantive and structural (Moreton & Pater, 2012a, 2012b). Both of these categories exclusively deal with individual processes. For example, a structural bias for featurally simple patterns has been demonstrated in phonotactics (see, e.g., Pycha et al., 2003), and a substantive bias for phonetically-grounded processes has been explored in phonological alternation learning (e.g. Wilson, 2006).

Few artificial language experiments have focused on biases involving the interaction of multiple processes, which are not easily categorized as structural or substantive. This paper explores two such interaction-based biases: *Maximal Utilization* (henceforth MaxUtil; Kiparsky, 1968), which favors patterns in which all rules are maximally utilized, and *Transparency* (Kiparsky, 1971), which favors interactions in which processes are not opaque. Differences in rule utilization are exemplified in (1), using a hypothetical interaction from Baković (2011).

(1) Transparent interactions

	Feeding	Bleeding
Underlying Representation (UR)	/tai/	/tia/
Deletion Rule ($V_1V_2 \rightarrow V_2$)	ti	ta
Palatalization Rule (t→tʃ/_i)	tʃi	-
Surface Representation (SR)	[tʃi]	[ta]

In Example (1), a deletion rule and a palatalization rule interact such that, depending on the UR that the grammar is given, either a feeding or bleeding interaction can occur. The underlying form /tai/ causes a feeding interaction, since the deletion rule creates the environment in which the other rule can apply. That is, when the vowel that triggers t→tʃ occurs second, deletion removes the intervening vowel, feeding palatalization. Conversely, the UR /tia/ illustrates a bleeding interaction, since the deletion rule removes the environment in which palatalization would otherwise take place (i.e. when the triggering vowel occurs first, deletion removes it before it can cause a change to the underlying /t/). Crucially, the existence of these interactions is dependent both on the rules' ordering (deleting, then palatalizing) and the data present in a language (either /tai/ or /tia/ in the underlying forms).

MaxUtil would favor the feeding interaction, since it utilizes both rules, while the bleeding interaction only utilizes one. Both the feeding and bleeding interactions from (1) would be favored by a Transparency Bias. This is because the underlying structures that would normally be removed or changed by the interactions' rules (i.e. a sequence of two vowels or a /ti/) are not found in any of their SR's, and since no structures that surface as a result of the rules appear outside those rules' relevant contexts. The opaque counterparts of these feeding and bleeding interactions are illustrated in (2), where the opposite rule ordering, paired with the same two UR's is shown:

(2) Opaque interactions

	Counterfeeding	Counterbleeding
UR	/tai/	/tia/
Palatalization Rule	-	tʃia
Deletion Rule	ti	tʃa
SR	[ti]	[tʃa]

The counterfeeding interaction shown in (2) is opaque, since a [ti] sequence is present in the SR, despite the palatalization rule being applied at one point in the derivation. This structure was able to survive because palatalization was applied too early to remove it. The counterbleeding interaction is also opaque, since a palatalized consonant is present outside of the environment *_i*. This happens because the /i/ that originally triggered the change was deleted after palatalization took place.

Kiparsky (1968, 1971) proposed the MaxUtil and Transparency Biases separately, based on historical evidence. Here I argue that they both affect language learners' ability to acquire phonological grammars, a proposal that was first explored computationally by Jarosz (2016a). To investigate this in humans using more direct evidence than Kiparsky (1968, 1971) had access to, I adapted the hypothetical interactions from (1) and (2) into toy languages, and used each in an artificial language learning experiment. I then ran simulations to test whether these biases can arise from the learning process (as suggested by Jarosz, 2016a, 2016b) or need to be a built-in aspect of the Grammar as Kiparsky assumed.

The structure of the paper is as follows: §2 discusses the theoretical work on biases favoring certain phonological interactions and reviews previous experimental and computational work on the subject, §3-5 discuss the design, methodology, and results of my experiment, which finds that both biases do exist in phonological learning when the relevant parts of the language are examined, §6 tests for these biases in two computational models and finds that they arise from the learning process, even in a model with almost no built-in linguistic structure, and §7 concludes.

2. Background

2.1. Theoretical work on interaction biases

Kiparsky (1968) proposed MaxUtil as a way of explaining certain trends that he observed in historical changes. For example, he noted that in Slavic languages two rules are ordered in a way that does not represent the chronological order in which they were introduced into the language (he treated this historical ordering as a default for speakers' synchronic grammars). Instead, the rules are ordered in a way that maximizes their utilization. This is shown in (3):

(3) Possible rule orderings in Slavic (transcriptions from Kiparsky 1968)

Feeding (actual ordering)		Counterfeeding (hypothetical)	
UR	/gělo/	UR	/gělo/
[+velar]→[+strident]	ʒělo	[ʒ]→[z]	-
[ʒ]→[z]	zělo	[+velar]→[+strident]	ʒělo
SR	[zělo]	SR	[ʒělo]

In the example above, the actual ordering in Slavic languages utilizes both the *[+velar]→[+strident]* rule, and as the *[ʒ]→[z] rule.* However, in the counterfeeding rule ordering, the interaction only utilizes the *[+velar]→[+strident]* rule when an underlying /gě/ sequence is present. Kiparsky (1968) argued that learners of Slavic imposed the feeding rule ordering because of a bias for maximal rule utilization.

Kiparsky (1968) also gave examples of a bias against bleeding interactions to support his MaxUtil hypothesis. One of these examples is shown in (4):

(4) Rule orderings in German dialects (transcriptions from Kiparsky 1968)

Counterbleeding (more common)		Bleeding (less common)	
UR	/tāg/	UR	/tāg/
Spirantization	tāɣ	Spirantization	tāk
Word-final devoicing	tāx	Word-final devoicing	-
SR	[tāx]	SR	[tāk]

The example above demonstrates two attested rule orderings in German: a counterbleeding order where spirantization applies before word-final devoicing, and a bleeding order in which the opposite is true. When devoicing occurs first, the grammar does not utilise spirantization, since it only affects [+voice] segments. However, if spirantization applies first, speakers utilise both rules. Since the counterbleeding ordering is more common among German dialects, this was further evidence for a MaxUtil Bias.

Taken together, these trends suggest that speakers generally preferred languages that maximized rules' "fullest utilization in the grammar" (Kiparsky, 1968, p. 200). However, this could not be the only factor affecting rule orderings, since bleeding orders also arose diachronically in a number of cases (Kenstowicz & Kisseberth, 1971). To address this, Kiparsky (1971) proposed a different factor governing rule ordering: a bias for orderings to be "maximally transparent" (Kiparsky, 1971, p. 623). Kiparsky defined a transparent rule ordering as one in which the conditions in (5) are met, while an opaque ordering was one in which at least one of these conditions was not true.

(5) For a rule with the structure A→B/C_D to be transparent (Kiparsky, 1971):

i. The SR must not contain a structure *B* which was created by the rule, but which no longer appears in the environment *C_D*…
ii. …Nor can the SR contain a structure *A* which appears in the context *C_D*, despite the rule having applied at one point in the derivation.

These conditions correspond to McCarthy's (1999) "surface apparent" and "surface true", respectively (see Baković, 2011 for more on this distinction). Counterbleeding violates the first condition for transparency, while counterfeeding violates the second condition. A Transparency Bias would therefore prefer feeding and bleeding orders, regardless of rule utilization.

2.2. Experimental work on opacity

In this section I will review the relatively small number of artificial language learning experiments that have investigated phonological interactions (Ettlinger, 2008; Kim, 2012; Brooks et al., 2013). For a detailed review of the artificial language learning literature in general, see Moreton and Pater (2012a, 2012b).

Ettlinger (2008) tested for the presence of both Transparency and MaxUtil Biases using a series of artificial language learning experiments. His artificial languages involved interacting vowel lowering and vowel harmony processes that were triggered by the addition of a suffix and prefix, respectively. These experiments found evidence that suggested a bias for transparent interactions, as well as a bias for faithful ones (i.e. bleeding and counterfeeding), which seemed to challenge Kiparsky's (1968) proposal for a MaxUtil Bias. However, there were issues with this conclusion. The bulk of Ettlinger's (2008) analysis involved individual t-tests on the accuracies associated with each type of stimulus to see if each condition's accuracy was significantly above chance. In addition to this making his results difficult to interpret, it led to a large number of inferential statistical tests (>30 t-tests for each experimental condition), and Ettlinger (2008) made no mention of correcting for multiple comparisons.

The other two experiments (Kim, 2012; Brooks et al., 2013) differed from Ettlinger's (2008) and the current study in that they withheld any evidence that would suggest a particular rule ordering to participants. Both of these studies were interested in which kinds of orderings would be chosen by default when participants were taught two rules that had the potential to interact. In testing, they exposed participants to forms that would force them to choose between an opaque and transparent rule ordering. Kim (2012) found a preference for counterfeeding over feeding orderings, however that study used no inferential statistics and a relatively small number of participants (N=12). Brooks et al. (2013) use a mixed-effects logistic regression to test their results and found that most of their participants chose to not utilise either rule when presented with a word in which both rules would interact. To my knowledge, this behavior is not attested in any natural language, and while it could be evidence for a bias toward faithful derivations, it could also be an experimental artifact.

In summary, while only a small number of experiments on interaction learning have been performed, the ones that exist challenge Kiparsky's (1968) original claims about MaxUtil. However, the majority of these experiments explored default rule orderings in the face of ambiguous training data, rather than testing how well participants could learn an ordering that was unambiguous. In the experiment presented in §3-5, learnability will be more directly explored to test for the existence of both MaxUtil and Transparency Biases.

2.3. Computational work on MaxUtil and Transparency Biases

When proposing a Transparency Bias, Kiparsky (1971) suggested that it would need to be built into the grammar. Specifically, he suggested that the "opacity of rules adds to the cost of the grammar[s]" that make use of them (Kiparsky, 1971, p. 614). He suggested that this cost could affect the learnability of interactions with opaque derivations and pressure speakers toward more transparent phonologies. This idea has been challenged recently, with computational studies showing that Transparency Bias can arise from the learning process without being formally added to a theory of acquisition.[1]

Jarosz (2016a) used an expectation-driven, Harmonic Serialism learner (henceforth the EDL model; Jarosz, 2015) with *Serial Markedness Reduction* (henceforth SMR; Jarosz, 2014) to simulate the acquisition of the four main interaction types: bleeding, feeding, counterbleeding, and counterfeeding. SMR uses *SM Constraints*, which enforce a particular ordering in which markedness constraints must be satisfied (cf. the "candidate chains" used by McCarthy, 2007). When these constraints evaluate each candidate, they operate on a list of all of the markedness constraints that candidate has satisfied up to that point in the derivation (these lists are shown inside of angle brackets below). By assigning violations to candidates that fail to satisfy markedness constraints in a particular order, SM Constraints are able to prefer opaque surface candidates to their transparent counterparts. A possible SMR analysis of the counterfeeding derivation from (2) is given below (for specific constraint definitions, see Jarosz, 2016a, pp. 2–3).

(6) Counterfeeding derivation using SMR

Step 1

/tai/	*VV	SM(*ti,*VV)	*ti	IDENT	MAX	SM(*VV,*ti)
tai<>	W*		L		L	
→ ti<*VV>			*		*	

[1] While several models exist that can learn opaque interactions (e.g. Rasin et al., 2017), in this section I present the only two that, to my knowledge, have been used to explore the relevant biases.

Step 2

ti<*VV>	*VV	SM(*ti,*VV)	*ti	IDENT	MAX	SM(*VV,*ti)
→ ti<*VV>			*			
tʃi<*VV, *ti>		W*	L	W*		

In the tableaux above, the UR /tai/ undergoes the deletion process in the first step of the derivation because *VV is ranked above MAX. No other changes can be made, since in Harmonic Serialism, GEN is restricted to making either 1 or 0 changes to the input per step. The deletion of /a/ satisfies *VV, adding this constraint to the winning candidate's list of satisfied markedness constraints. The candidate *ti<*VV>* is then passed as input into the second step, where SM(*ti,*VV) causes the faithful candidate to win. This constraint assigns violations to any form that satisfies *VV and *ti in either the order <*VV, *ti> or simultaneously. Since the form *tʃi<*VV, *ti>* meets the former criterion, it violates SM(*ti, *VV). This means that the ranking of SM(*ti,*VV) over *ti is crucial to attain a counterfeeding derivation. Otherwise, in the second step, the winning candidate would always be the one that palatalizes and the process would be a feeding derivation instead of a counterfeeding one.

The EDL model acquires rankings like those in (6) using expectation-driven learning. To do this, the learner's initial state assigns equal probability to every possible ranking of constraints. It then gradually assigns more probability to the correct rankings by seeing which increase its ability to predict the training data. To see which rankings provide better predictions, the model steps through each pair of constraints one-by-one, and samples from the pair's two possible rankings. The sampling process assumes a probability of 1 for the ranking that's currently being tested and samples from all the other rankings using the model's current estimation of how likely they are. The probability of a ranking is increased if it produces a relatively high number of samples that matched the training data, and decreased if it fails to produce many matching samples. For more on this process, see Jarosz (2015) or the software's documentation at https://github.com/gajajarosz/hidden-structure.

Jarosz (2016a) tested which biases arise from the EDL model with SMR by measuring the number of iterations it took to converge on each type of interaction, given different kinds of learning data. She found that when the training data included interacting forms 10 times as often as other forms, the model showed a MaxUtil Bias. However, when interacting forms were 10 times less frequent than other forms, the EDL model showed a Transparency Bias. Data that was not skewed in either direction caused the model to be relatively unbiased. These results support the idea that MaxUtil and Transparency Biases could arise from skewed data (Jarosz, 2016a, 2016b), but suggests that they may not be active when learning data is evenly distributed. However, in §6.1, I present novel results using the EDL model that show both biases arise in its learning, even when balanced training data is used.

Nazarov and Pater (2017) used a maximum entropy implementation of Stratal OT (Kiparsky, 2000) to see if a Transparency Bias would emerge over the course of learning in a stratal framework. Their model was trained with L-BFGS-B (Byrd et al., 1995), an optimization algorithm that updates weights based on an objective function that the algorithm attempts to minimize. Their metric for learnability was how reliably the model converged on each pattern. While they found some evidence for a bias toward transparent derivations (i.e. their model converged more often when trained on these languages than it did on opaque ones), this disappeared when the model was given more realistic learning data. They did not test for any effects of MaxUtil on the model's learning. Since the probability of convergence is not directly related to the experimental results presented here, I leave further exploration of this model's biases to future work.

3. Design

3.1. Artificial Languages

To test whether MaxUtil and Transparency Biases are present in human learning, I adapted hypothetical examples from Baković (2011) for bleeding, feeding, counterbleeding, and counterfeeding languages with one language representing each interaction type. The stems in the artificial languages ended in either /t/, /d/, /k/, or /g/. There were two suffixes in each language: /-i/ and /-a/. These affixes could occur individually (in which case they marked either the diminutive or plural) or together (in which case they marked the diminutive-plural). When cooccurring, the relative order of the affixes determined whether the language was bleeding/counterbleeding (/i+a/) or feeding/counterfeeding (/a+i/). Which of the suffixes corresponded to diminutive and which corresponded to plural was counterbalanced across conditions. Two rules were present in the language: a palatalization rule and a vowel deletion rule. Example (6) shows both rules in standard notation.

(6) Rules in the artificial language

Palatalization	[Coronal] → [-anterior]/_[+high]	(i.e. [t, d] → [ʧ, ʤ]/_[i])
Vowel Deletion	[+Syll.] → Ø / _[+Syll.]	(i.e. V_1V_2 → V_2)

The ordering of these rules determined if a language was transparent or opaque. If the deletion rule was ordered before the palatalization rule, the derivation was transparent, and if the palatalization rule was ordered first, the derivation was opaque.

The different possible combinations of word-final segments and affixation created four different trial types for participants: trials in which forms faithfully surfaced, trials where only deletion applied, trials where only palatalization applied, and trials where deletion and palatalization interacted. These are shown for each of the four languages in Table 1. In the table, only stem-final consonants and affixes are shown, and voiceless consonants represent themselves and their voiced counterparts.

Table 1. Trial Types for Each Language

	Faithful[2]	*Deleting*	*Palatalizing*	*Interacting*
Bleeding				
/UR/	/t+a/	/k+i+a/	/t+i/	/t+i+a/
Deletion	-	ka	-	ta
Palatalization	-	-	ʧi	-
[SR]	[ta]	[ka]	[ʧi]	[ta]
Feeding				
/UR/	/t+a/	/k+a+i/	/t+i/	/t+a+i/
Deletion	-	ki	-	ti
Palatalization	-	-	ʧi	ʧi
[SR]	[ta]	[ki]	[ʧi]	[ʧi]
Counterbleeding				
/UR/	/t+a/	/k+i+a/	/t+i/	/t+i+a/
Palatalization	-	-	ʧi	ʧia
Deletion	-	ka	-	ʧa
[SR]	[ta]	[ka]	[ʧi]	[ʧa]
Counterfeeding				
/UR/	/t+a/	/k+a+i/	/t+i/	/t+a+i/
Palatalization	-	-	ʧi	-
Deletion	-	ki	-	ti
[SR]	[ta]	[ki]	[ʧi]	[ti]

[2] Trials involving stimuli with UR's containing /k+a/ and /k+i/ were also categorized as Faithful.

3.2. Predictions

If only MaxUtil Bias is present in human learning, feeding and counterbleeding should be the most learnable languages. However, if Transparency is the only bias present, bleeding and feeding should be more learnable than their counterparts. Furthermore, while past work has focused on how these biases' affect either interacting items alone (e.g. Brooks et al., 2013) or the acquisition of the language as a whole (e.g. Jarosz, 2016a), these are not the only possible outcomes. Since the utilization and transparency of the deletion rule stays constant across all languages and trial types, one should expect the acquisition of that rule to be unaffected by the two biases. On the other hand, since evidence for the palatalization rule *is* affected by the interaction type, acquisition of that pattern should be affected by these biases for both interacting and palatalizing items. For these reasons, rather than expecting language-wide effects of MaxUtil and Transparency biases, I expect them to primarily affect participant accuracy in palatalizing and interacting trial types.

4. Methodology

4.1. Stimuli

The stimuli used in the experiment included all 72 possible stems of the shape {i, u}{m, n, l}{i, a, u}{t, k, d, g} and all 216 possible combinations of these stems with the three different affixation patterns (/-i/, /-a/, or both). These stimuli were presented auditorily and were recorded in a sound-attenuated booth by a native English speaker (the author) using an M-Audio Fast Track Pro Mobile Audio Interface, a Shure SM10A head-worn microphone, and the software Audacity (sample rate and bit depth were left at their default settings of 44100 Hz and 32-bit, respectively).

Visual stimuli communicated semantic information in the experiment. Eight image categories contained nine images each: fruits/vegetables, animals, apparel, transportation, body parts, natural structures (e.g. trees, volcanos, etc.), man-made structures, and tools. These 72 images were randomly paired with the 72 stems for each participant. The experiment software communicated plural forms by presenting two copies of an image and illustrated diminutives with a smaller version of the image. For diminutive-plurals, the software presented two smaller copies of the image (see Ettlinger, 2008 for a similar presentation style for diminutive and plural items).

4.2. Participants

Participants (N=48) were recruited using Mechanical Turk (https://www.mturk.com/) and paid $4.75 for their participation (UMass Amherst IRB protocol number 2017-4040). There were 23 participants who identified as male, 23 who identified as female, and 2 who chose to not answer in regards to their sex. Three participants did not report their age, but for those who did, the mean age was 34.87 (SD=9.10). Participants were all self-reported native speakers of English and they reported a number of second languages such as Spanish, Japanese, and Dutch.

4.3. Procedure

The experiment was advertised on Mechanical Turk using the title "Learn an alien language," the description "Receive training on an alien language and answer questions about what you've learned. Native English speakers only. Please don't redo this HIT if you've done it in the past," and the key words "alien, language, game, learning, linguistics." After choosing to take part in the experiment, participants were taken to a page where they could read an informed consent form as well as instructions. These instructions are given in (7).

(7) Experiment Instructions

In this experiment, you will be asked to learn aspects of an imaginary "alien" language. The experiment should take about 30 minutes.

- First, you'll answer three questions about English to practice using the experiment software. Once you've answered these correctly, you'll move on to the Training Section.
- In the Training Section you'll be asked questions about the alien language, and you'll receive feedback on your answers. It's okay to guess at first, since you'll be learning by trial and error. When you've finished this section, feel free to take a break.
- Languages often have rules that apply when a suffix is added to a word. For example, the 'f' in 'hoof' changes to a 'v' in the plural form ('hooves'). Try to figure out rules like this that are at work in the alien language.[3]
- The last section will be the Testing Section. It will be just like the Training Section, but you will no longer receive feedback on your answers.

Please be sure to wear headphones while participating and do not take notes of any kind during the Training Section. Feel free to contact the researcher at [AUTHOR'S EMAIL ADDRESS] if you have any questions regarding the experiment. Press BEGIN to start the practice questions.

Once participants clicked the first "Begin" button below these instructions, they were taken to a set of practice questions. These were meant to ensure that they understood how to use the experiment software, that they could clearly hear the audio stimuli, and that they understood the learning task they were supposed to perform. The first question presented participants with pictures of a dog and a cat, one of which was circled. Beneath the pictures, the software presented the question "What's the word for the circled item(s)?", as well as two audio files that played automatically. One of the files was the word "dog" said aloud and the other was the word "cat." Participants had to click a button that corresponded to the audio file that matched the circled picture (i.e. perform a forced-choice task) and were immediately given feedback that told them whether or not their choice was correct. Then, they were taken to the second practice question: this was similar to the first, but showed a picture depicting a single lock and one showing a pair of locks (the latter of which was circled). Two audio clips played the words "lock" and "locks" and participants had to choose which corresponded to the circled picture. They again received feedback and were taken to the next practice question. The third question was similar to the first two, although the images showed a full-sized statue and a smaller statue (the latter of which was circled). The recordings played the words "statue" and "statuette." Participants had to again choose the recording that matched the circled item.

If participants answered incorrectly for any of the practice questions, they were reminded of the instructions and had to answer all three of the questions again until they were able to do them correctly. If they answered them all correctly, they were brought to a page that said, "Nice work! You answered all the practice questions correctly. Press BEGIN to start the experiment" and could begin the experiment by clicking a button labeled "Begin."

Each experiment trial first presented participants with a single, full-sized picture and automatically played an audio recording of the stem that corresponded to that picture. Above the recording was text saying "Word for this Item:". After the recording of the stem finished, participants were automatically taken to a page that was similar to the practice questions. Two images were presented: one that was identical to the

[3] Pilot results suggested that this instruction was necessary for participants to reach a reasonable level of accuracy in testing. While instructions like this have been shown to encourage explicit learning in artificial language experiments (Moreton & Pertsova, 2016), I leave exploring the possible effect of these instructions to future work (see §5.2 for more on what strategies participants reported using in training).

singular, full-sized image they were just shown and one that was either smaller, doubled, or both (the latter picture had a circle around it). Two audio files played below the images and participants were instructed to choose which audio file correctly represented the circled picture. Depending on the trial type, the two audio files could represent one of three forced-choice tasks: choosing between a palatalized and non-palatalized stem-final consonant, choosing between deleting or not deleting the first vowel in a VV sequence, or choosing between the correct and incorrect ordering of rules. The latter task appeared on all of the interacting trials, the first task appeared in palatalizing and faithful trials[4], and the second task appeared in deleting trials. Table 2 demonstrates the options that participants were choosing between in each trial type and language. The circled image always appeared on the right side of the screen, but which side the correct audio file appeared on was randomized. Once participants chose an audio file, the trial ended.

The experiment contained two phases: training and testing. In training, participants received feedback in the form of either "Correct!" or "Incorrect." at the end of each trial. In testing, no feedback was given. All three affixed forms for 36 of the stems appeared in training and all three forms for the remaining 36 appeared in testing. Within each phase, the trial in which each form appeared was randomized. Since the stems in training and testing were completely different, participants had to use a generalized process in testing, rather than memorizing how individual words behaved.

Table 2. Choices Presented to Participants in Each Trial Type and Language

	Faithful	*Deleting*	*Palatalizing*	*Interacting*
Bleeding				
/UR/	/t+a/	/k+i+a/	/t+i/	/t+i+a/
[correct SR] (incorrect SR)	[ta] (tʃa)	[ka] (kia)	[tʃi] (ti)	[ta] (tʃa)
Feeding				
/UR/	/t+a/	/k+a+i/	/t+i/	/t+a+i/
[correct SR] (incorrect SR)	[ta] (tʃa)	[ki] (kai)	[tʃi] (ti)	[tʃi] (ti)
Counterbleeding				
/UR/	/t+a/	/k+i+a/	/t+i/	/t+i+a/
[correct SR] (incorrect SR)	[ta] (tʃa)	[ka] (kia)	[tʃi] (ti)	[tʃa] (ta)
Counterfeeding				
/UR/	/t+a/	/k+a+i/	/t+i/	/t+a+i/
[correct SR] (incorrect SR)	[ta] (tʃa)	[ki] (kai)	[tʃi] (ti)	[ti] (tʃi)

At the end of the experiment, participants answered a small number of demographic and debriefing questions, such as "How did you approach the learning task?" and "When were you born?".

5. Results

5.1. Accuracy in the Testing Phase

As predicted in §3.2, no language-wide effects of either bias were found to be significant. Figure 1 illustrates participants' accuracy in the testing phase, organized by trial type and then by language. When tested on interacting forms, participants in the bleeding and feeding conditions performed best, with participants in the two opaque languages doing significantly worse. This is the pattern predicted by a Transparency Bias, since bleeding and feeding both meet the conditions in (5). This is evidence that a Transparency Bias could affect how easily the ordering of rules is acquired.

[4] In the case of stems with /k/ and /g/ as final consonants, their palatalized forms ended in [tʃ] and [dʒ], respectively. Choosing an SR with a palatalized form was never the correct choice for these words.

For palatalizing trials, bleeding and counterfeeding participants performed at chance, while feeding and counterbleeding participants performed significantly better. This is the patterning predicted by a MaxUtil Bias, since these two languages both utilise the palatalization rule in every relevant derivation. This suggests that MaxUtil could specifically affect the learning of whatever rule's utilization is language-dependent (i.e. palatalization for the languages tested here).

Trials involving faithful forms (i.e. underlying /…ta/, /…ki/, and /…ka/) showed no major differences across languages. This is not surprising, since the faithful trials were identical across all conditions. That is, participants were all trained to faithfully map /…ta/, /…ki/, and /…ka/, regardless of which language they were learning. The relatively low accuracy across conditions for faithful trials could be the result of a the experiment instructions specifically guiding participants to learn unfaithful alternations in the training phase.

Counterfeeding was favored over counterbleeding in deleting trials (i.e. underlying /…kai/ or /…kia/, depending on the language). This isn't predicted by either bias but could be the result of some confounding factors present in English. While none of the stimuli were English words, some of them did end in English words. Participants could have been biased toward choosing surface forms that did this when their alternative was a form with a less familiar ending. Specifically, the correct choice for deleting forms in the counterfeeding language was […ki], while the incorrect choice for deleting forms in the counterbleeding language was […kia]. These are both English words, with [ki] referring to a device used to unlock doors and [kia] referring to a car company. Their alternatives ([…kai] and […ka], respectively) are not words in English, which could have biased participants toward choosing the more English-like forms. Furthermore, words ending in [ki], [gi], and [kia] are more frequent in English than words ending in [kai], [gai], [ka], or [ga], which could have also contributed to this confound (frequencies were obtained from CELEX; Baayen et al., 1995). Fortunately, this issue was not present in any of the other trial types, and the deleting trials were not one of the two trial types of interest.

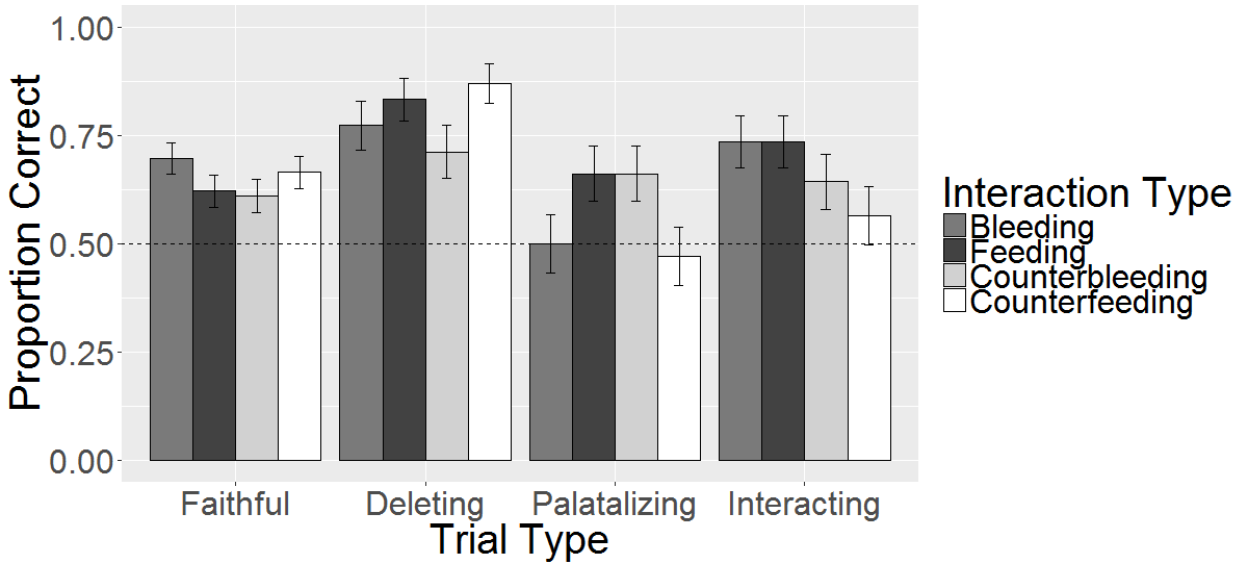

Figure 1. Average percent correct in testing, by pattern type and trial type. See Table 1 for explanations of each trial type. Error bars show 95% confidence intervals calculated over all datapoints in each condition.

Additionally, deleting forms showed a higher overall accuracy than the other trial types. This is likely due to the fact that the deletion rule was simpler than the palatalization one. Participants learned to delete all vowel-vowel sequences, but had to learn to palatalise a more specific set of the consonant-vowel sequences they were exposed to (see Moreton & Pater, 2012a for more on this kind of complexity in phonological learning). Since it likely took time for them to acquire the correct context for palatalization (i.e. /ti/ and /di/), and since faithful forms could be incorrectly palatalized in the case where that context had not yet been learned, the higher accuracy for deleting forms is expected.

To test for the statistical significance of these observations, I ran a mixed effects logistic regression with each bias and each rule as a separate variable. The coding for these variables is shown in (8) and (9). If a MaxUtil Bias exists, then participants who learned feeding and counterbleeding languages should have higher performance in the palatalizating trials than those who were trained on bleeding and counterfeeding patterns. This would result in an estimate for MaxUtil*Palatalization in the model that is significantly higher than zero. If a Transparency Bias is present, participants in the feeding and bleeding conditions should have higher accuracy in interacting trials than those who were learning counterbleeding and counterfeeding languages. This would produce an estimate for Transparent*Deletion*Palatalization that is significantly higher than zero. Likewise, if Transparency and MaxUtil Biases do not affect participants' learning, then MaxUtil*Palatalization and Transparent*Deletion*Palatalization, respectively, should not significantly differ from zero.

(8) Coding of Bias Variables

Language	Transparent	MaxUtil
Bleeding	1	0
Feeding	1	1
Counterbleeding	0	1
Counterfeeding	0	0

(9) Coding of Rule Variables

Trial Type	Palatalization	Deletion
Faithful	0	0
Deleting	0	1
Palatalizing	1	0
Interacting	1	1

The model was run using the *lme4* (Bates et al., 2015) package in R (R Core Team, 2016), with random effects for participant and stimulus on the intercept. The interactions of interest, MaxUtil*Palatalization ($z=4.480$, $p<.001$) and Transparent*Deletion*Palatalization ($z=2.377$, $p<0.02$), both had significantly positive estimates. These support the interpretation of Figure 1 above, in which both the MaxUtil and Transparency Biases seem to each be affecting one of the relevant parts of the language. Results from the full model are given in the Appendix.

5.2. Debriefing responses

Participants answered a number of debriefing questions describing their experience in the experiment, as well as their demographic information. For a summary of participant demographics, see §4.2. There were two questions regarding participants' experiences that yielded meaningful answers. The first asked them to grade how well they thought they did in the testing phase. The majority of participants (N=30) thought that they did better than chance (>50%), with a mean grade of 55.89 (SD=18.99). This shows that on average, participants were aware that they were acquiring the language's phonology (which is true, considering the overall mean accuracies by language). The other relevant information gained from participants' answers in

this section of the experiment was that a majority of participants (N=25) reported answering in the test phase based on intuition, and a majority (N=35) also reported using a rule or pattern (these were not mutually exclusive choices for them to report). This suggests that participants were not approaching the experiment as a memorization task.

6. Learning simulations

The results in §5 show that MaxUtil Bias and Transparency Bias *do* affect human learning, even when participants' learning data is not skewed toward interacting or non-interacting items. However, these biases were only apparent in their relevant trial types. Jarosz (2016a) found that her EDL model showed no major, language-wide biases when trained on unbiased data, however she did not break down her model's performance any further. That is, her conclusions were based on the model's average performance across all kinds of trials. In this section, I first report results from simulations of my experiment using the Jarosz (2016a) model to see if its learning, when analyzed in a way similar to my experiment results, shows the same kinds of biases as the human participants. Then, to see if the EDL model's biases are dependent on the linguistic structure built into it, I show results from a neural network with a minimal amount of prespecified linguistic structure.

6.1. Harmonic Serialism with SMR Constraints

I gave the EDL model learning data that matched the words and frequencies present in the experimental training data (i.e. 18 of each unique word ending), in the form of UR→SR mappings. These frequencies were roughly equivalent to the UNI condition tested by Jarosz (2016a). The constraints I used are those given in the SMR analysis from §2.3, with the learner's sample-size parameter set to 1000 and the training mode set to "batch". The GEN function for the model was given the ability to delete the first vowel in VV sequences, and to palatalise both velar and alveolar stops. Since Harmonic Serialism only allows GEN to make one or zero changes in a single step, this meant that the model's choice of candidate in any given tableau was relatively small, as illustrated in (12).

(12) Candidates input type for the EDL Model (following Jarosz 2016a)

Input	Candidates
tia	{tia, ta, tʃia}
tai	{tai, ti, tʃai}
ti	{ti, tʃi}
ta	{ta, tʃa}
kia	{kia, ka, tʃia}
kai	{kai, ki, tʃai}
ki	{ki, tʃi}
ka	{ka, tʃa}
tʃia	{tʃia, tʃa}
tʃai	{tʃai, tʃi}
tʃi	{tʃi}
tʃa	{tʃa}

After running the model in 20 separate simulations, I estimated its average accuracy at each iteration. This estimate was obtained by sampling from the grammar 100 times for each UR. I used this sample to estimate the probabilities of each SR, given the UR that had been used as input. I used these UR→SR probabilities to calculate the probability of an accurate response in each of the experiment's trial types, using the equation in (13), following Luce (1959):

(13) Accuracy estimation (EDL Model)

$$Accuracy = \frac{\Pr(correct\ form)}{\Pr(correct\ form) + \Pr(incorrect\ form)}$$

For example, if the model was trained on the feeding language and then estimated probabilities of .8, .1, and .1 for the mappings /tai/→[tʃi], /tai/→[tai], and /tai/→[ti], respectively, the model's accuracy would be the probability of the correct form (.8) divided by the sum of the probabilities for both of the forms that participants chose between (.8 for [tʃi] + .1 for [ti]). So the final accuracy measurement would be .8/(.8+.1)≈.889. Crucially, the probability of /tai/→[tai] is left out of the calculation, since participants were not able to choose this SR in the experiment. If one of the forms that participants were choosing between was not in the model's final list of possible output candidates, it was assigned a probability of zero. This calculation was performed separately for each run, and the average accuracy at each iteration, for the two relevant trial types in each language, is shown in Figures 2 and 3.

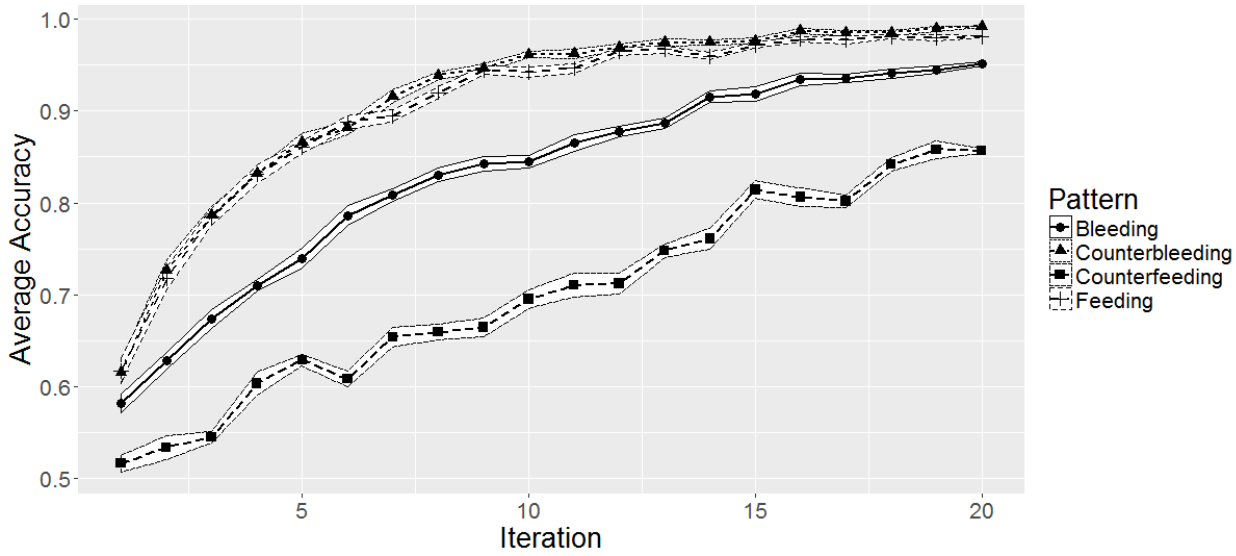

Figure 2. Average accuracy in palatalizing trials for the EDL model across iterations in a forced-choice task. Results are broken down by pattern type. White area behind the lines shows standard error of the mean, calculated over all datapoints in each condition.

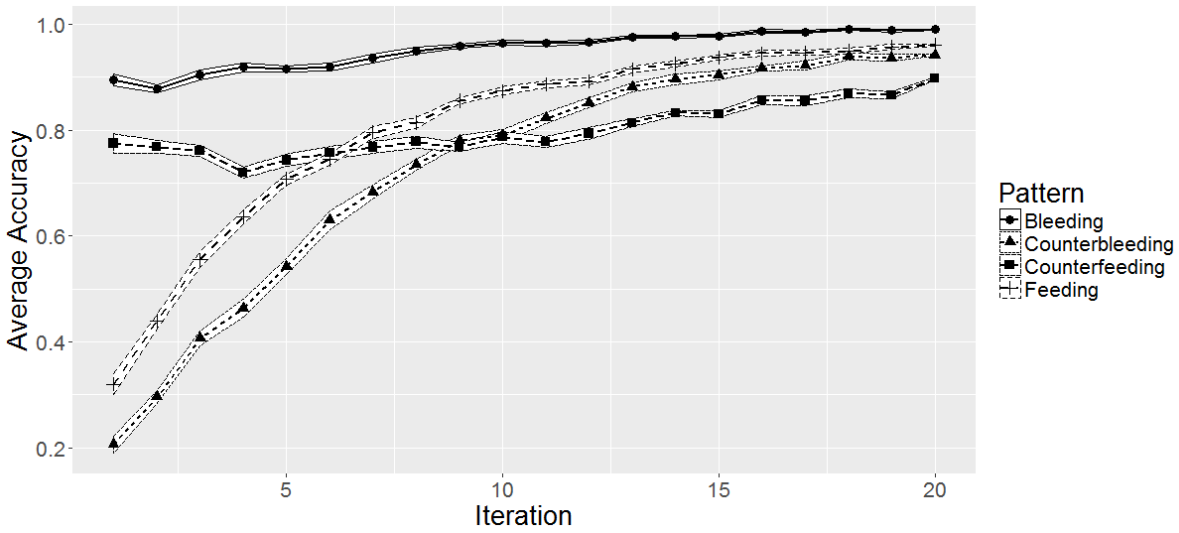

Figure 3. Average accuracy in interacting trials for the EDL model across iterations in a forced-choice task. Results are broken down by pattern type. White area behind the lines shows standard error of the mean, calculated over all datapoints in each condition.

While some variation exists across these learning curves (e.g. the feeding language takes some time to become more accurate than the opaque languages in interacting trials), the relative ordering of the patterns is stable throughout most of the learning process. To more closely compare the model's results with those that were observed in humans, the accuracy estimates from the end of learning (shown in Figure 4 in a format that's comparable to Figure 1) were analyzed using two-way, independent samples t-tests. These confirmed that the model had significantly higher accuracies for transparent languages (i.e. bleeding and feeding) in interacting trials ($t[78] = 28.994$, $p < .0001$) and maximally utilizing languages (i.e. feeding and counterbleeding) in palatalizing trials ($t[78] = 9.5468$, $p < .0001$). This demonstrates that when the learner's results are broken down by trial type, it has both Transparency and MaxUtil Biases in the relevant parts of the language, even when the learning data is not skewed in the way that Jarosz's (2016a) was. These trends directly mirror the experiment results[5] presented in §5.1.

This demonstrates that these biases do not need to be externally added to the model, as Kiparsky (1971) originally suggested. Instead, the biases arise from the learning process itself. Specifically, a MaxUtil Bias emerges in the model's accuracy for palatalizing forms, because to accurately produce those forms, the model must correctly rank the *ti constraint over IDENT. In the feeding and counterbleeding languages, palatalizing and interacting forms both provide unambiguous data to support this ranking. In counterfeeding languages, interacting forms provide evidence that's ambiguous between the ranking IDENT>>*ti and SM(*ti,*VV)>>*ti>>IDENT. Bleeding languages have similarly ambiguous data—with interacting forms supporting rankings of IDENT>>*ti as well as the correct ranking of *ti>> IDENT. This ambiguity in the model's training data causes it to take longer to learn that particular ranking, lowering its average accuracy on palatalizing trials.

A Transparency Bias arises for interacting forms because less evidence exists for the ranking of SM constraints than for the ranking of the markedness and faithfulness constraints. Since palatalizing, faithful,

[5] Unlike the humans, the EDL model achieves perfect performance on faithful trials. This is the result of the relatively limited constraint set causing some non-faithful candidates to be harmonically bound, and would likely be different if the constraint set were more exhaustive.

and deleting forms don't require the SM constraints to be correctly ranked, a Transparency Bias does not exist for these forms. However, for the model to correctly produce interacting forms in the counterbleeding and counterfeeding languages, the crucial SM constraints must be ranked above IDENT and *ti, respectively. Since the interacting forms are the only evidence the model has for these SM constraints' correct rankings, it takes longer for the model find them (see Jarosz, 2016a for a similar discussion of the EDL model's difficulty in learning opaque interactions).

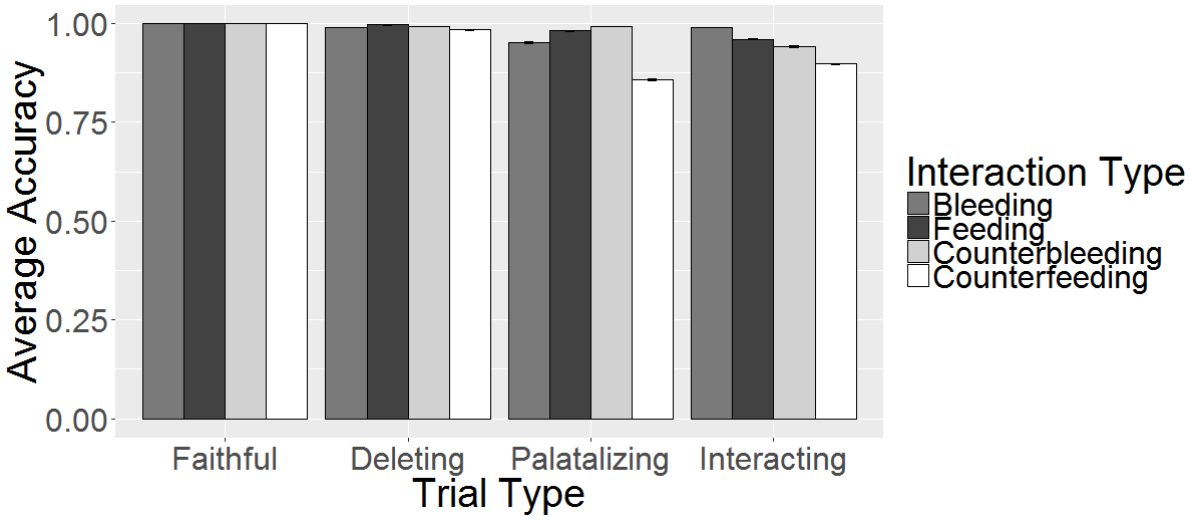

Figure 4. Average accuracy for the EDL model at the end of learning, broken down by pattern type and trial type. Error bars show standard error of the mean, calculated over all datapoints in each condition. A lack of error bars represents SEM measurements that are too small to be seen in the figure.

6.2. Sequence-to-Sequence Neural Network

The results from the EDL model showed that Transparency and MaxUtil don't have to be explicitly added to the cost of a grammar to correctly reproduce human-like learning. This section examines whether this result will be true for a model that has less linguistic structure explicitly worked into its learning process.

I used a Sequence-to-Sequence neural network (henceforth Seq2Seq; Sutskever et al., 2014), implemented using the *Seq2Seq* package (Rahman, 2016) in Python. This neural network architecture was originally designed for machine translation, but has been proposed as a baseline for morphophonological learning and correlates well with human behavior in a number of such tasks (Kirov, 2017). For example, when tested on the experimental results from Albright and Hayes (2003), a Seq2Seq model's predictions correlated with human behavior more than any previously proposed model (Kirov & Cotterell, 2018; although see Corkery et al., 2019 for a critique of these results). The Seq2Seq network learns string-to-string mappings (UR to SR mappings, in this case) by updating weights for connections between nodes that are organized into multiple layers. It includes no explicit linguistic structure in its architecture, unlike the EDL model which includes hand-crafted phonological constraints, as well as assumptions about how those constraints interact with one another (e.g. that learning consists of figuring out which constraints strictly dominate others). Since this structure is absent, the Seq2Seq's learning task involves not only discovering which patterns exist in a language and how they interact, but also how to best represent those patterns using its connection weights.

For these simulations, the only explicit linguistic structure that the Seq2Seq model was given was a set of standard phonological features, shown in Table 3. Since the model uses real-numbered values in its input and output, I used 1, -1, and 0 to represent +'s, -'s, and unmarked feature values, respectively.

Table 3. Phonological Features Given to Seq2Seq Model

Segment	[syllabic]	[voice]	[Coronal]	[anterior]	[Dorsal]	[high]
[t]	-	-	+	+		
[d]	-	+	+	+		
[tʃ]	-	-	+	-		
[dʒ]	-	+	+	-		
[k]	-	-			+	
[g]	-	+			+	
[i]	+	+				+
[a]	+	+				-

Words were represented only by their endings (i.e. the final consonant and everything occurring after it), which were encoded as a list of segments, where each segment was a vector of phonological feature values. Input and output length were both set to 3 segments (since the model requires these lengths be a constant value throughout the simulation, words of length 2 ended in an empty symbol that was unmarked for every feature). Frequencies in the training data matched those present in the training phase of the experiment, and the model was run for 20 iterations in each simulation, with a batch size equal to the total number of training data. I used RMSProp (a standard, error-based learning algorithm used for neural networks; Tieleman & Hinton, 2012) to minimise the model's mean squared error (MSE), a total of 4 hidden layers (with 15 nodes in each layer), and a learning rate of 0.005. MSE was calculated by going through each feature in the model's predicted output, squaring the difference between the predicted value of this feature and the correct value, and averaging across all of these squared differences.

At each iteration in training, the model was tested in a way that was similar to the forced-choice procedure used by participants and simulated by the EDL model. However, since the Seq2Seq network is not probabilistic, the model's MSE for incorrect and correct forms was used to estimate accuracy. For example, the model would be given a UR, such as /ti/, and then its MSE was calculated for both the correct SR, [tʃi], and the incorrect[6] choice that participants were presented with in the experiment, [ti]. If the model learned the correct UR→SR mappings, its MSE for the incorrect form would be higher than the MSE for the correct form. The equation that I used to convert these error metrics into an accuracy score comparable to the one used for the EDL model is shown in (14):[7]

(14) Accuracy estimation (Seq2Seq Model)

$$Accuracy = \frac{\text{MSE}(incorrect\ form)}{\text{MSE}(correct\ form) + \text{MSE}(incorrect\ form)}$$

I used this as an estimate of the model's accuracy across each trial type in each language. Figures 5 and 6 show the average accuracy estimates over 20 separate simulations for each iteration of learning. As in §6.1, the relative ordering of languages is stable across much of the model's learning in both the interacting and palatalizing contexts (with the exception of the feeding language, which again took some time to become

[6] Technically the MSE calculation is only "error" in the case of correct forms (since the model's output *should* be different than the incorrect forms). In this case, the MSE measure is used for convenience, but using Euclidean distance would give equivalent results.

[7] For a similar approach that uses network error to estimate probability, see Kurtz (2007).

more accurate than its opaque counterparts on interacting items). And like the EDL model, transparent languages tend to have higher accuracies in the interacting trials, while maximally utilizing languages tend to be more accurate in palatalizing trials. To confirm these qualitative results, I again performed two-way t-tests on the model's results from the end of learning (see Figure 7 for a presentation of these results that's comparable to Figure 1). These showed that both transparent languages ($t[78] = 2.452$, $p = 0.016$) and maximally utilizing languages ($t[78] = 6.98$, $p < .0001$) had significantly higher accuracies in the relevant trial types.

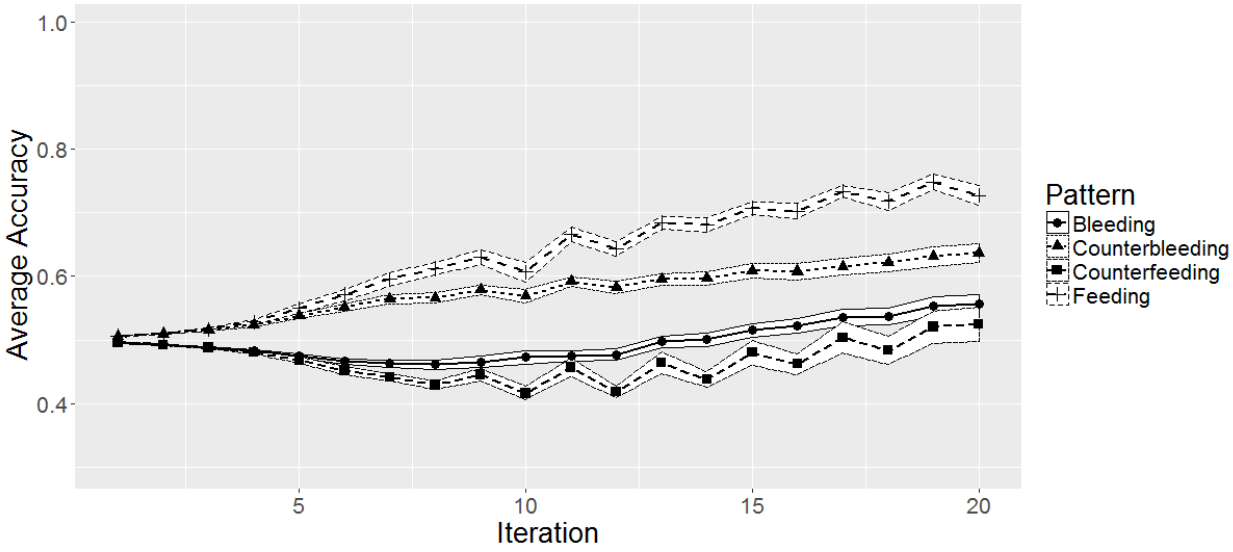

Figure 5. Average accuracy in palatalizing trials for the Seq2Seq model across iterations in a forced-choice task. Results are broken down by pattern type. White area behind the lines shows standard error of the mean, calculated over all datapoints in each condition.

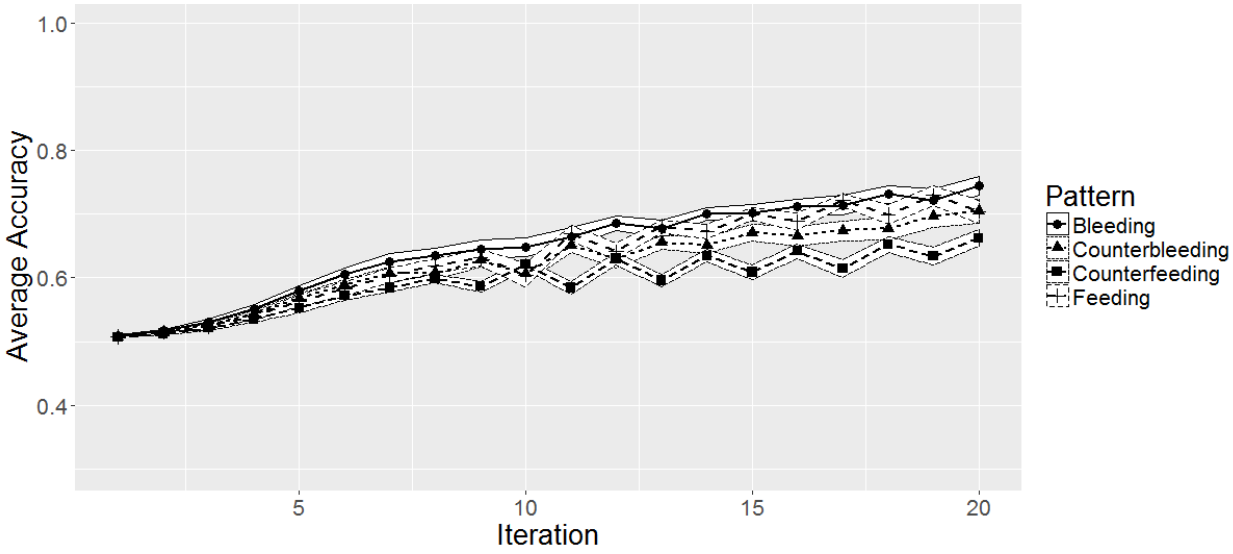

Figure 6. Average accuracy in interacting trials for the Seq2Seq model across iterations in a forced-choice task. Results are broken down by pattern type. White area behind the lines shows standard error of the mean, calculated over all datapoints in each condition.

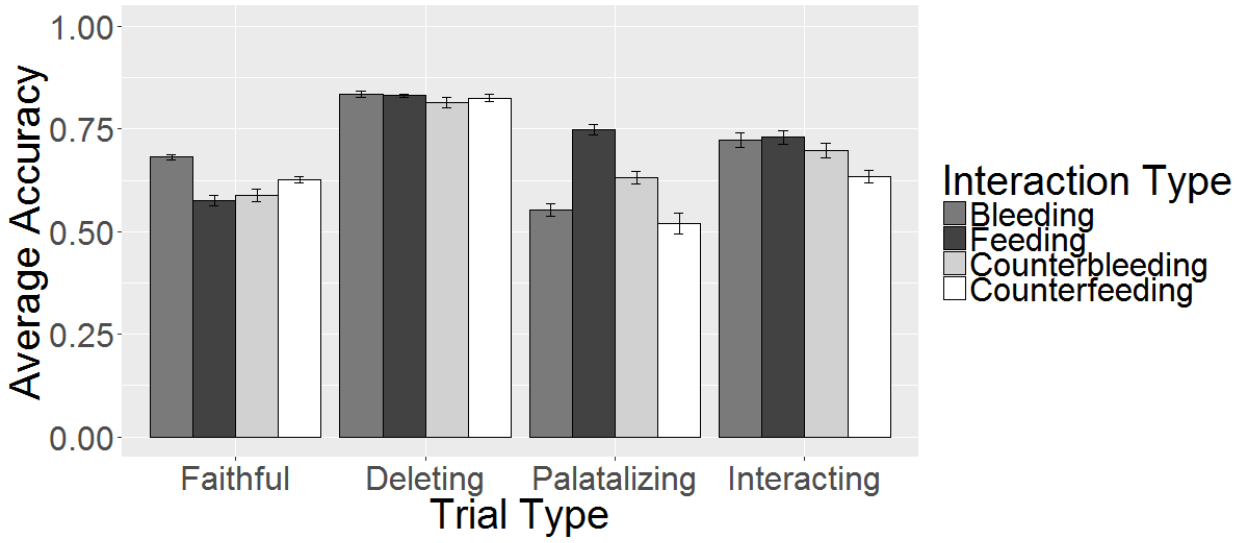

Figure 7. Average accuracy for the Seq2Seq model at the end of learning, broken down by pattern type and trial type. Error bars show standard error of the mean, calculated over all datapoints in each condition.

These results suggest that even without the built-in linguistic structure used by the EDL model, the two biases of interest emerge is the Seq2Seq network's learning. Understanding this learner's biases is more difficult than in the simpler EDL model. While there are likely some similarities in the cause for a MaxUtil Bias (i.e. ambiguous data causing the model to learn more slowly), the presence of a Transparency Bias is less easy to explain. A number of methods are being developed to better understand the representations that neural networks use when learning linguistic patterns (for example, see Li et al., 2015; Giulianelli et al., 2018; McCoy et al., 2018), however this an ongoing area of research and attempting to perfectly understand the Seq2Seq model's biases would be outside of the scope of this paper.

The conclusion that *can* be drawn from these results is that the kind of explicit, linguistic structure present in the EDL model is not necessary to simulate the MaxUtil and Transparency Biases. Both of Kiparsky's (1968, 1971) biases emerge in the Seq2Seq model's learning, despite the lack of hand-made constraints or a GEN component controlling what kinds of mappings the model can produce. This suggests that these biases are not dependent on the EDL learner's assumptions and would likely emerge in a variety of phonological learning models.

7. Discussion

7.1. Future Work

Participants could have analyzed the patterns they were learning in a number of other ways. For example, they could have learned the differences across suffix and stem SR's as allomorphy or phonological exceptionality (for proposals along these lines, see Sanders, 2003). Since participants did not achieve perfect accuracy, a number of imperfect solutions could have also been reached (such as a palatalization rule that applies to all /t/'s in the counterbleeding language). Because this study was concerned with testing whether Kiparsky's (1968, 1971) proposed biases occurred in human learning (regardless of the analysis that those humans used), figuring out the exact pattern that participants learned is outside of the scope of this paper. However, future work could explore this topic more carefully by testing participants'

generalization to novel suffixes, and by using procedures more fine-grained than a forced-choice between two surface forms (e.g. the kind of production task used by Brooks et al., 2013).

A number of other questions remain for future work: for example, both the experiment participants and the computational models discussed here were given information about the UR's for the forms they were trained on. However, UR-learning could be an important factor in real-world interaction acquisition, and future research could examine its effects in both experimental and computational domains. Other avenues of research include directly modeling the diachronic changes Kiparsky observed (see Zuraw, 2003 for more on simulating historical change) and investigating rule interactions other than the four discussed here, such as Duke-of-York derivations (Pullum, 1976; see Baković, 2011 for a extensive review of different interaction types).

Finally, other models with the capability for learning opaque patterns should be tested, to see how well they predict the behavior observed here. Nazarov and Pater's (2017) Stratal MaxEnt model, Rasin et al.'s (2017) minimum description length learner, and general-purpose learning algorithms for finite state automata (Chandlee & Jardine, 2014) should all be able to learn the patterns that the experiment participants were trained on. However, future work is needed to understand how to best compare the learning in these frameworks to human data.

7.2. Conclusions

The experiment results presented in this paper support the existence of both of Kiparsky's (1968, 1971) biases in phonological learning. However, I found that each bias only affects a portion of the language-learning process. While a Transparency Bias exists in the acquisition of particular rule orderings, this bias did not affect the learning of the rules themselves. Instead, a MaxUtil Bias was the predominant factor that affected how easily participants learned the palatalization rule in the experiment. This localized effect of the MaxUtil Bias could be why Brooks et al. (2013) found what seemed to be the opposite pressure in their experiment. Since their participants never saw a rule being underutilized in training, the MaxUtil Bias observed here could not have affected their participants' learning.

The results in this paper also helped to shed light on the inherent biases present in Jarosz's (2016a) model. While she found no language-wide biases in the EDL model when it was given balanced training data, the simulations presented in §6.1 showed that biases were apparent when the model's performance was broken down by trial type. When this was done, the EDL model showed the same pattern of MaxUtil and Transparency Biases as the participants in my experiment. This suggests that these biases need not be explicitly built into a theory of grammar as Kiparsky (1971) suggested. In fact, a model with less explicit linguistic structure than the EDL model was able to model both biases successfully. This suggests that MaxUtil and Transparency Biases can both arise from the learning process itself and are not dependent on a model being given explicit, linguistic structure.

In summary, this paper showed that in an artificial language learning experiment, participants demonstrated a bias for maximal utilization of rules, as well as a bias for transparent rule orderings. While these biases were only present in particular parts of the language, this is the first piece of direct evidence that they could both play a role in shaping the diachronic changes observed by Kiparsky (1968, 1971). Furthermore, I showed that these biases can arise without being explicitly added into a model of phonological learning—challenging Kiparsky's (1971) assumption to the contrary.

References

Albright, Adam, & Hayes, Bruce. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Baayen, R. Harald, Piepenbrock, Richard, & Gulikers, Leon. (1995). The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

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

Bates, Douglas, Maechler, Martin, Bolker, Ben, Walker, Steven, Christensen, Rune Haubo Bojesen, Singmann, Henrik, … Grothendieck, Gabor. (2015). Package 'lme4.' *Convergence*, *12*(1).

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

Byrd, Richard H., Lu, Peihuang, Nocedal, Jorge, & Zhu, Ciyou. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Chandlee, Jane, & Jardine, Adam. (2014). Learning phonological mappings by learning Strictly Local functions. *Proceedings of the Annual Meetings on Phonology*, *1*.

Corkery, Maria, Matusevych, Yevgen, & Goldwater, Sharon. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. *ArXiv Preprint ArXiv:1906.01280*.

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

Giulianelli, Mario, Harding, Jack, Mohnert, Florian, Hupkes, Dieuwke, & Zuidema, Willem. (2018). Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information. *ArXiv Preprint ArXiv:1808.08079*.

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

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

Jarosz, Gaja. (2016a). Learning Opaque and Transparent Interactions in Harmonic Serialism. *Proceedings of the Annual Meetings on Phonology*, *3*.

Jarosz, Gaja. (2016b). *Refining UG: Connecting Phonological Theory and Learning*. Presented at the North East Linguistics Society, University of Massachusetts Amherst. Retrieved from https://blogs.umass.edu/jarosz/files/2016/10/NELS2016_final_red.pdf

Kenstowicz, Michael J., & Kisseberth, Charles W. (1971). *Unmarked bleeding orders*.

Kim, Yun Jung. (2012). Do learners prefer transparent rule ordering? An artificial language learning study. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, *48*, 375–386. Chicago Linguistic Society.

Kiparsky, Paul. (1968). *Linguistic universals and linguistic change*.

Kiparsky, Paul. (1971). *Historical linguistics*.

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

Kirov, Christo. (2017). Recurrent Neural Networks as a Strong Domain-General Baseline for Morpho-Phonological Learning. *Poster Presented at the 2017 Meeting of the Linguistic Society of America*.

Kirov, Christo, & Cotterell, Ryan. (2018). *Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker & Prince (1988) and the Past Tense Debate*.

Kurtz, Kenneth J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Li, Jiwei, Chen, Xinlei, Hovy, Eduard, & Jurafsky, Dan. (2015). Visualizing and understanding neural models in NLP. *ArXiv Preprint ArXiv:1506.01066*.

Luce, R. Duncan. (1959). *Individual choice behavior.*

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

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

McCoy, R. Thomas, Linzen, Tal, & Frank, Robert. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *ArXiv Preprint ArXiv:1802.09091*.

Moreton, Elliott, & Pater, Joe. (2012a). Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass*, *6*(11), 686–701.
Moreton, Elliott, & Pater, Joe. (2012b). Structure and substance in artificial-phonology learning, part II: Substance. *Language and Linguistics Compass*, *6*(11), 702–718.
Moreton, Elliott, & Pertsova, Katya. (2016). Implicit and explicit processes in phonotactic learning. *Proceedings of the 40th Boston University Conference on Language Development, Somerville, Mass., Pp. TBA. Cascadilla*.
Nazarov, Aleksei, & Pater, Joe. (2017). Learning opacity in Stratal Maximum Entropy Grammar. *Phonology*, *34*(2), 299–324.
Pullum, Geoffrey K. (1976). The Duke of York gambit. *Journal of Linguistics*, *12*(1), 83–102.
Pycha, Anne, Nowak, Pawel, Shin, Eurie, & Shosted, Ryan. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. *WCCFL*, *22*, 423–435. Linguistics Department, Stanford University.
R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Retrieved from http://www.R-project.org/
Rahman, Fariz. (2016). *seq2seq: Sequence to Sequence Learning with Keras* [Python]. Retrieved from https://github.com/farizrahman4u/seq2seq
Rasin, Ezer, Berger, Iddo, Lan, Nur, & Katzir, Roni. (2017). *Rule-based learning of phonological optionality and opacity*. Presented at the North East Linguistic Society. Retrieved from http://www.mit.edu/~rasin/files/abstracts/nels2017.pdf
Sanders, Robert Nathaniel. (2003). *Opacity and sound change in the Polish lexicon* (PhD Thesis).
Sutskever, Ilya, Vinyals, Oriol, & Le, Quoc V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104–3112.
Tieleman, Tijmen, & Hinton, Geoffrey. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, *4*(2), 26–31.
Wilson, Colin. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945–982.
Zuraw, Kie. (2003). Probability in language change. *Probabilistic Linguistics*, 139–176.

Appendix

The full mixed effects model used in §5.1 is given in Example (i). The coefficients that are relevant to the MaxUtil and Transparency Biases are shown in italics:

(i) Logistic regression with participant and item as random effects on the intercept:

Predictor	Estimate	Std. Error	z-value	p-value
(Intercept)	0.86485	0.10670	8.106	5.25e-16 ***
Transparent	0.11605	0.10367	1.119	0.26296
MaxUtil	0.02974	0.10367	0.287	0.77423
Deletion	0.33472	0.04429	7.557	4.12e-14 ***
Palatalization	-0.28054	0.04431	-6.332	2.43e-10 ***
Transparent*MaxUtil	0.04931	0.10490	0.470	0.63830
Transparent*Deletion	0.06648	0.03692	1.801	0.07173
MaxUtil*Deletion	-0.08276	0.03693	-2.241	0.02502 *
Transparent*Palatalization	0.07296	0.03692	1.976	0.04813 *
*MaxUtil*Palatalization*	*0.20447*	*0.03693*	*5.536*	*3.09e-08* ***
Deletion*Palatalization	-0.09566	0.04419	-2.165	0.03041 *
Transparent*MaxUtil*Deletion	0.07837	0.04025	1.947	0.05155
Transparent*MaxUtil*Palatalization	-0.10818	0.04025	-2.688	0.00720 **
*Transparent*Deletion*Palatalization*	*0.08761*	*0.03686*	*2.377*	*0.01747* *
MaxUtil*Deletion*Palatalization	-0.07654	0.03687	-2.076	0.03789 *
Transp*MaxUtil*Del*Pal	-0.11767	0.04021	-2.927	0.00343 **