Phonology **36** (2019) 627–653. © Cambridge University Press 2020
doi:10.1017/S0952675719000320

*Learning biases in opaque interactions**

Brandon Prickett
University of Massachusetts Amherst

This study uses an artificial language learning experiment and computational modelling to test Kiparsky's claims about Maximal Utilisation and Transparency biases in phonological acquisition. A Maximal Utilisation bias would prefer phonological patterns in which all rules are maximally utilised, and a Transparency bias would prefer patterns that are not opaque. Results from the experiment suggest that these biases affect the learnability of specific parts of a language, with Maximal Utilisation affecting the acquisition of individual rules, and Transparency affecting the acquisition of rule orderings. Two models were used to simulate the experiment: an expectation-driven Harmonic Serialism learner and a sequence-to-sequence neural network. The results from these simulations show that both models' learning is affected by these biases, suggesting that the biases emerge from the learning process rather than any explicit structure built into the model.

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, b). 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 categorised as

* E-mail: BPRICKETT@UMASS.EDU.
Thanks to Eric Baković, as well as the attendees of NECPhon 2017 and LabPhon 2018 for helpful discussion about the topics in this paper. The members of UMass's Sound Workshop and Phonology Reading Group, as well as the anonymous reviewers, also provided helpful insight. Most of all, I owe a thank you to Gaja Jarosz and Joe Pater for guidance, discussion and assistance throughout the research process. This study was supported by NSF grants #BCS-1650957 and #BCS-424077.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

structural or substantive. This paper explores two such interaction-based biases: MAXIMAL UTILISATION (henceforth MaxUtil; Kiparsky 1968), which favours patterns in which all rules are maximally utilised, and TRANSPARENCY (Kiparsky 1971), which favours interactions in which processes are not opaque. Differences in rule utilisation are exemplified in (1), using a hypothetical interaction adapted from Baković (2011).

(1) *Transparent interactions*

	feeding	*bleeding*
Underlying Representation (UR)	/tai/	/tia/
Deletion ($V_1V_2 \rightarrow V_2$)	ti	ta
Palatalisation (t → ʧ/__i)	ʧi	—
Surface Representation (SR)	[ʧi]	[ta]

In (1), a deletion rule and a palatalisation 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/ leads to 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/ → [ʧ] occurs in V_2 position, deletion removes the intervening vowel, feeding palatalisation. Conversely, the UR /tia/ illustrates a bleeding interaction, since the deletion rule removes the environment in which palatalisation 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 ordering of the rules (deletion, then palatalisation) and the data present in a language (either /tai/ or /tia/ in the underlying forms).

MaxUtil would favour the feeding interaction, since it utilises both rules, while the bleeding interaction only utilises one. Both the feeding and bleeding interactions from (1) would be favoured by a Transparency bias. This is because neither interaction results in an opaque relationship between underlying and surface forms. That is, there are no structures in the SR that would normally be removed or changed by the interactions' rules and there are no changes made by the rules that surface outside of their environment. The opaque counterparts of these feeding and bleeding interactions are illustrated in (2), where the opposite rule ordering is shown.

(2) *Opaque interactions*

	counterfeeding	*counterbleeding*
UR	/tai/	/tia/
Palatalisation	—	ʧia
Deletion	ti	ʧa
SR	[ti]	[ʧa]

The counterfeeding interaction shown in (2) is opaque, since a [ti] sequence is present in the SR, despite the palatalisation rule having

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

applied in the derivation. This structure is able to survive because palatalisation applies too early to remove it. The counterbleeding interaction is also opaque, since a palatalised consonant is present outside of the environment __[i], because the /i/ that originally triggers the change is deleted after palatalisation took place.

Kiparsky (1968, 1971) proposes 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 incorporated 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, b), or need to be a built-in aspect of the grammar, as assumed by Kiparsky.

The structure of the paper is as follows: §2 discusses the theoretical work on biases favouring 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) proposes MaxUtil as a way of explaining certain trends in historical changes. For example, he notes (1968: 197–198) that in some Slavic languages the two rules in (3) are ordered in a way that does not represent the chronological order in which they were introduced (he treated this historical ordering as a default for speakers' synchronic grammars). Instead, the rules are ordered in a way that maximises their utilisation.

(3) a. *Feeding*

UR	/gelo/
[g] → [ʤ] / __ [e, i]	ʤelo
[ʤ] → [z]	zelo
SR	[zelo]

b. *Counterfeeding*

UR	/gelo/
[ʒ] → [z]	—
[g] → [ʤ] / __ [e, i]	ʤelo
SR	[ʤelo]

In the example above, the Slavic languages with a feeding interaction utilise both [g] → [ʤ] and [ʒ] → [z]. However, in the counterfeeding languages, the interaction only utilises [g] → [ʤ] when an underlying /ge/ sequence is present. Kiparsky (1968) argues the Slavic language learners

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

who imposed a feeding rule ordering must have been pressured to do so by a bias for maximal rule utilisation.

Kiparsky also gives examples of a bias against bleeding interactions in support of his MaxUtil hypothesis. One example is shown in (4) (1968: 199).

(4) a. *Counterbleeding* (*more common*)

UR	/taːg/
Spirantisation	taːɣ
Word-final devoicing	taːx
SR	[taːx]

b. *Bleeding* (*less common*)

UR	/taːg/
Word-final devoicing	taːk
Spirantisation	—
SR	[taːk]

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

Taken together, these trends suggest that speakers generally prefer languages that maximise rules' 'fullest utilization in the grammar' (Kiparsky 1968: 200). However, this might not be the only factor affecting rule orderings, since bleeding orders also arise diachronically in a number of cases (Kenstowicz & Kisseberth 1971). To address this, Kiparsky (1971: 623) proposes a different factor governing rule ordering: a bias for orderings to be 'maximally transparent'. He defines a transparent rule ordering as one in which the conditions in (5) are met, while an opaque ordering is one in which at least one of these conditions does not hold.

(5) For a rule with the structure $A \rightarrow B / C__D$ to be transparent:
   a. The SR must not contain a structure B which was created by the rule, but which no longer appears in the environment C__D …
   b. … 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 utilisation.

2.2 Experimental work on opacity

In this section I will review the three artificial language learning experiments that have investigated phonological interactions (Ettlinger 2008,

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

Kim 2012, Brooks *et al*. 2013). For a detailed review of the artificial language learning literature in general, see Moreton & Pater (2012a, b).

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), apparently challenging Kiparsky's (1968) proposal for a MaxUtil bias. However, there are issues with this conclusion. The bulk of Ettlinger's (2008) analysis involves 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 leads to a large number of inferential statistical tests (>30 *t*-tests for each experimental condition), and Ettlinger (2008) makes no mention of correcting for multiple comparisons.

The other two experiments (Kim 2012, Brooks *et al*. 2013) differ from Ettlinger (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 a transparent rule ordering. Kim (2012) found a preference for counterfeeding over feeding orderings; however, the study used no inferential statistics and a relatively small number of participants ($N = 12$). Brooks *et al*. (2013) used 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 behaviour 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 artefact.

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 unambiguous ordering. 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

Kiparsky (1971) proposes that a Transparency bias would need to be built into the grammar. Specifically, he suggests that 'opacity of rules adds to the cost of a grammar' that makes use of them (1971: 623). The cost could affect the learnability of interactions with opaque derivations, and

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

pressure speakers toward more transparent phonologies. This idea has been challenged recently, with computational studies showing that the Transparency bias can arise from the learning process without being formally added to a theory of acquisition.[1]

Jarosz (2016a) uses an expectation-driven Harmonic Serialism learner (Jarosz 2015; henceforth the EDL model) with Serial Markedness Reduction (Jarosz 2014; henceforth SMR) 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 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 in (6) (for specific constraint definitions, see Jarosz 2016a: 2–3).

(6) a. *Step 1*

/tai/	*VV	SM(*ti,*VV)	*ti	IDENT	MAX	SM(*VV,*ti)
i. tai ⟨⟩	W*		L		L	
☞ ii. ti ⟨*VV⟩			*		*	

b. *Step 2*

ti ⟨*VV⟩	*VV	SM(*ti,*VV)	*ti	IDENT	MAX	SM(*VV,*ti)
☞ i. ti ⟨*VV⟩			*			
ii. ʧi ⟨*VV,*ti⟩		W*	L	W*		

In the tableaux in (6), the UR /tai/ undergoes deletion 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 one or zero changes to the input at every 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 to 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 either in the order ⟨*VV, *ti⟩ or simultaneously. Since the form [ʧi ⟨*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

[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.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

candidate would always be the one that palatalises, 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 grammar by testing different possible rankings, and rewarding those that 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 is 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 matches the training data, and decreased if it fails to produce many matching samples.[2]

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 ten times as often as other forms (her 'HI' condition), the model showed a MaxUtil bias. However, when interacting forms were ten times less frequent than other forms, the EDL model showed a Transparency bias. Data that was not skewed in either direction (her 'UNI' condition) caused the model to be relatively unbiased. These results support the idea that MaxUtil and Transparency biases could arise from skewed data (Jarosz 2016a, b), but suggests that they may not be active when learning data is evenly distributed. However, in §6.1, using the EDL model, I present novel results that show that both biases arise in its learning, even when balanced training data is used.

Nazarov & Pater (2017) use 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 optimisation algorithm that updates weights based on an objective function that the algorithm attempts to minimise. Their metric for learnability was the reliability of the model's convergence 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 when trained 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.

[2] For more on this process, see Jarosz (2015) or the software documentation at https://github.com/gajajarosz/hidden-structure.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

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 the affixes co-occurred, their relative order determined whether the language was bleeding/counterbleeding (/-i + -a/) or feeding/counterfeeding (/-a + -i/). Which of the suffixes corresponded to diminutive and which to plural was counterbalanced across conditions. Two rules were present in the language: a palatalisation rule and a deletion rule. The rules are shown in (7).

(7) *Rules in the artificial language*

Palatalisation	[t d] → [ʧ ʤ] / __ [i]
Deletion	$V_1V_2 \rightarrow V_2$

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

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

3.2 Predictions

If only the 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 utilisation and transparency of the deletion rule stays constant across all languages and trial types, we should expect the acquisition of that rule to be unaffected by the two biases. On the other hand, since evidence for the palatalisation rule *is* affected by the interaction type, acquisition of that pattern should be affected by these biases for both interacting and palatalising

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

		faithful	deleting	palatalising	interacting
bleeding	UR	/t-a/	/k-i-a/	/t-i/	/t-i-a/
	Deletion	—	ka	—	ta
	Palatalisation	—	—	ʧi	—
	SR	[ta]	[ka]	[ʧi]	[ta]
feeding	UR	/t-a/	/k-a-i/	/t-i/	/t-a-i/
	Deletion	—	ki	—	ti
	Palatalisation	—	—	ʧi	ʧi
	SR	[ta]	[ki]	[ʧi]	[ʧi]
counterbleeding	UR	/t-a/	/k-i-a/	/t-i/	/t-i-a/
	Palatalisation	—	—	ʧi	ʧia
	Deletion	—	ka	—	ʧa
	SR	[ta]	[ka]	[ʧi]	[ʧa]
counterfeeding	UR	/t-a/	/k-a-i/	/t-i/	/t-a-i/
	Palatalisation	—	—	ʧi	—
	Deletion	—	ki	—	ti
	SR	[ta]	[ki]	[ʧi]	[ti]

Table I
Trial types for each language. Trials involving stimuli with URs containing /k-a/ and /k-i/ were categorised as faithful.

items. For these reasons, rather than expecting language-wide effects of the MaxUtil and Transparency biases, I anticipate that they will primarily affect participant accuracy in palatalising 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). The stimuli 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 Audacity software (sample rate and bit depth were left at their default settings of 44,100 Hz and 32 bit respectively). All stimuli were presented to participants auditorily as .wav files.

Visual stimuli communicated semantic information in the experiment. There were eight image categories, each containing nine images: fruits/vegetables, animals, apparel, transportation, body parts, natural structures (e.g. trees, volcanoes, etc.), man-made structures and tools. These 72 images were randomly paired with the 72 stems for each participant.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

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 of 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 and 23 as female, and two 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; 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 reproduced in (8). 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 one picture depicting a single lock and another 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'

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

and 'statuette'. Participants again had to choose the recording that matched the circled item.

(8) *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 E-MAIL ADDRESS] if you have any questions regarding the experiment. Press BEGIN to start the practice questions.

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 labelled '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 full-sized image of the singular that they had just been shown and one that was either smaller, doubled or both (the latter picture had a circle around it). Two audio files played automatically, and participants were instructed to choose which audio file correctly represented the circled

[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).

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

		faithful	deleting	palatalising	interacting
bleeding	UR	/t-a/	/k-i-a/	/t-i/	/t-i-a/
	correct SR	[ta]	[ka]	[ʧi]	[ta]
	incorrect SR	[ʧa]	[kia]	[ti]	[ʧa]
feeding	UR	/t-a/	/k-a-i/	/t-i/	/t-a-i/
	correct SR	[ta]	[ki]	[ʧi]	[ʧi]
	incorrect SR	[ʧa]	[kai]	[ti]	[ti]
counterbleeding	UR	/t-a/	/k-i-a/	/t-i/	/t-i-a/
	correct SR	[ta]	[ka]	[ʧi]	[ʧa]
	incorrect SR	[ʧa]	[kia]	[ti]	[ta]
counterfeeding	UR	/t-a/	/k-a-i/	/t-i/	/t-a-i/
	correct SR	[ta]	[ki]	[ʧi]	[ti]
	incorrect SR	[ʧa]	[kai]	[ti]	[ʧi]

Table II
Choices presented to participants in each trial type and language.

picture. Depending on the trial type, the two audio files could represent one of three forced-choice tasks: a choice between a palatalised and non-palatalised stem-final consonant, a choice between deleting and not deleting the first vowel in a VV sequence or a choice between the correct and incorrect ordering of rules. The latter task appeared in all of the interacting trials, the first task appeared in palatalising and faithful trials[4] and the second task appeared in deleting trials. Table II shows the options 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 randomised. Once participants had chosen 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 randomised. Since the stems in training and testing were completely different, participants had to use a generalised procedure in testing, rather than memorising how individual words behaved.

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?'.

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

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

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, organised by trial type and 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.

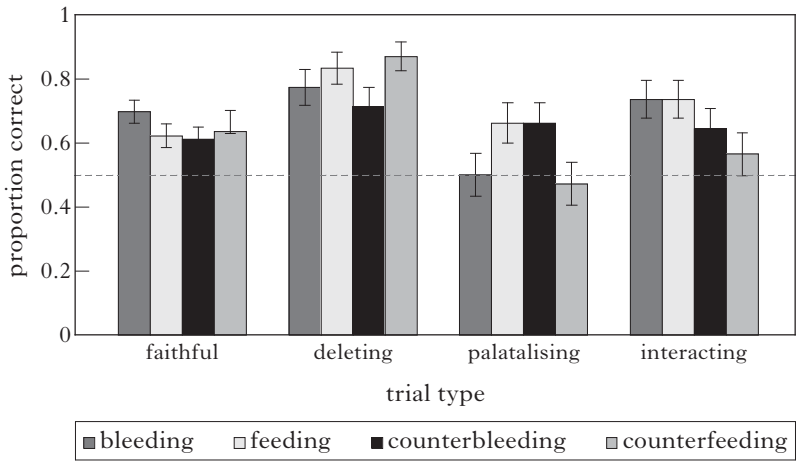


Figure 1

Average proportion of correct answers in the testing phase, by pattern type and trial type. See Table I for explanations of each trial type. Error bars show 95% confidence intervals calculated over all datapoints in each condition.

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

Trials involving faithful forms (i.e. underlying /…ta, …ki, …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, …ka/, regardless of which language they were learning. The relatively low accuracy across conditions for faithful trials could be the result of the experiment instructions

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

(a)

Language	Transparent	MaxUtil
bleeding	1	0
feeding	1	1
counterbleeding	0	1
counterfeeding	0	0

(b)

Trial type	Palatalisation	Deletion
faithful	0	0
deleting	0	1
palatalising	1	0
interacting	1	1

Table III
Coding of (a) bias variables; (b) rule variables.

specifically guiding participants to learn unfaithful alternations in the training phase.

Counterfeeding was favoured 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 sequences which could be English words. Participants could have been biased toward choosing such a surface form when the 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 […ki.a]. These are both English words, *key* and *Kia* respectively. Their alternatives ([…ka.i] and […ka] respectively) are not words in English, which could have biased participants toward choosing the more English-like forms. Fortunately, this issue was not present in any of the other trial types, and deletion was not one of the two trial types of interest.

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 palatalisation 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 probably took time for them to acquire the correct context for palatalisation (i.e. /ti, di/), and since faithful forms could be incorrectly palatalised 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 separate

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

predictor	estimate	SE	*z*	*p*
(Intercept)	0.865	0.107	8.106	5.25e-16***
Transparent	0.116	0.104	1.119	0.263
MaxUtil	0.030	0.104	0.287	0.774
Deletion	0.335	0.044	7.557	4.12e-14***
Palatalisation	−0.281	0.044	−6.332	2.43e-10***
Transparent × MaxUtil	0.049	0.105	0.470	0.638
Transparent × Deletion	0.066	0.037	1.801	0.072
MaxUtil × Deletion	−0.083	0.037	−2.241	0.025*
Transparent × Palatalisation	0.073	0.037	1.976	0.048*
MaxUtil × Palatalisation	*0.204*	*0.037*	*5.536*	*3.09e-08****
Deletion × Palatalisation	−0.096	0.044	−2.165	0.030*
Transparent × MaxUtil × Deletion	0.078	0.040	1.947	0.052
Transparent × MaxUtil × Palatalisation	−0.108	0.040	−2.688	0.007**
Transparent × Deletion × Palatalisation	*0.088*	*0.037*	*2.377*	*0.017**
MaxUtil × Deletion × Palatalisation	−0.077	0.037	−2.076	0.038*
Transparent × MaxUtil × Deletion × Palatalisation	−0.118	0.040	−2.927	0.003**

Table IV

Logistic regression with participant and item as random effects on the intercept. Coefficients relevant to the MaxUtil and Transparency biases are given in italics.

variables. The coding for these variables is shown in Table III. If a MaxUtil bias exists, then participants who learned feeding and counterbleeding languages should have higher performance in the palatalising trials than those who were trained on bleeding and counterfeeding patterns. This would result in an estimate for MaxUtil × Palatalisation 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 × Palatalisation that is significantly higher than zero. Likewise, if Transparency and MaxUtil biases do not affect participants' learning, then MaxUtil × Palatalisation and Transparent × Deletion × Palatalisation, respectively, should not significantly differ from zero.

The model was run using the *lme4* package (Bates *et al*. 2015) in R (R Core Team 2016), with random effects for participant and stimulus on the intercept. The interactions of interest, MaxUtil × Palatalisation ($z = 4.480$, $p < 0.001$) and Transparent × Deletion × Palatalisation ($z = 2.377$, $p < 0.02$), both had significantly positive estimates. These support the interpretation of Fig. 1 above, in which both the MaxUtil and

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

Transparency biases seem to each be affecting one of the relevant parts of the language. Results from the full model are given in Table IV.

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 reflected in the overall mean accuracies by language). The other relevant information gained from participants' answers in this section of the experiment was that a majority ($N = 25$) 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). This suggests that participants were not approaching the experiment as a memorisation task.

6 Learning simulations

The results in §5 show that the MaxUtil and Transparency biases *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 examine whether its learning, when analysed in a way similar to my experiment results, shows the same kinds of biases as the human participants. Then, to establish whether 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.[5]

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

[5] For the software used in these simulations and for the experiment data described in §5 see the online supplementary materials at https://doi.org/10.1017/S0952675719000320.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

constraints used are those given in the SMR analysis in §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 (9).

(9) *Candidates for each input type for the EDL model*

input	*candidates*	*input*	*candidates*
tia	[tia, ta, ʧia]	ki	[ki, ʧi]
tai	[tai, ti, ʧai]	ka	[ka, ʧa]
ti	[ti, ʧi]	ʧia	[ʧia, ʧa]
ta	[ta, ʧa]	ʧai	[ʧai, ʧi]
kia	[kia, ka, ʧia]	ʧi	[ʧi]
kai	[kai, ki, ʧai]	ʧa	[ʧa]

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. These UR → SR probabilities were then used to calculate the probability of an accurate response in each of the experiment's trial types, using the equation in (10), following Luce (1959).

(10) *Accuracy estimation* (*EDL model*)

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

For example, if the model that was trained on the feeding language estimated probabilities of 0.8, 0.1, and 0.1 for the mappings /tai/ → [ʧi], /tai/ → [tai] and /tai/ → [ti] respectively, its accuracy would be the probability of the correct form (0.8) divided by the sum of the probabilities for both of the forms that participants chose between (0.8 for [ʧi] + 0.1 for [ti]). So the final accuracy measurement would be $0.8/(0.8 + 0.1) = 0.889$. 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; the average accuracy at each iteration for the two relevant trial types in each language is shown in Fig. 2.

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

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

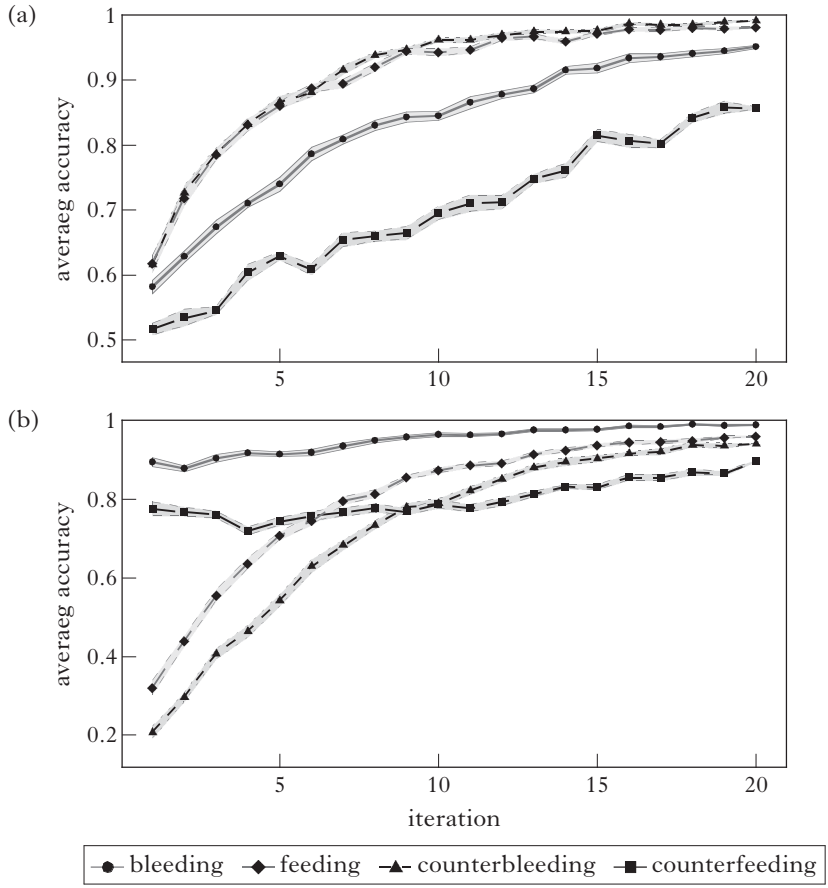


Figure 2

Average accuracy in (a) palatalising trials and (b) interacting trials for the EDL model across iterations in a forced-choice task. Results are broken down by pattern type. The grey area behind the lines shows standard error of the mean, calculated over all datapoints in each condition.

model's results with those that were observed in participants, the accuracy estimates from the end of learning (shown in Fig. 3 in a format comparable to Fig. 1) were analysed using two-way independent sample *t*-tests. These confirmed that the model had significantly higher accuracies for transparent languages (i.e. bleeding and feeding) in interacting trials ($t[78] = 28.994$, $p < 0.0001$) and maximally utilising languages (i.e. feeding and counterbleeding) in palatalising trials ($t[78] = 9.547$, $p < 0.0001$). 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

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

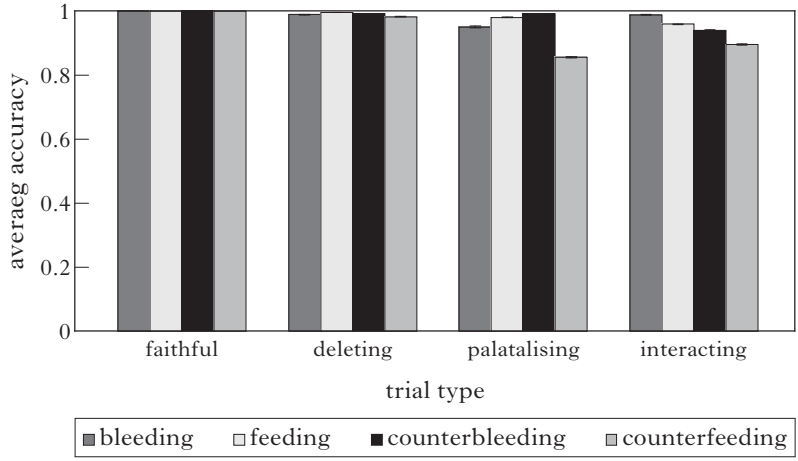


Figure 3

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 standard error of the mean measurements that are too small to be seen in the figure.

that Jarosz' (2016a) data was. These trends directly mirror the experiment results presented in §5.1.[6]

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 palatalising forms, because, to accurately produce those forms, the model must correctly rank the *ti constraint over IDENT. In the feeding and counterbleeding languages, palatalising and interacting forms both provide unambiguous data to support this ranking. In counterfeeding languages, interacting forms provide evidence that is ambiguous between the ranking IDENT ≫ *ti and SM(*ti,*VV) ≫ *ti ≫ IDENT. Bleeding languages have similarly ambiguous data – with interacting forms supporting the ranking IDENT ≫ *ti as well as the correct *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 palatalising trials.

A Transparency bias is found for interacting forms, because less evidence exists for the ranking of SM constraints than for the ranking of the markedness and faithfulness constraints. Since palatalising,

[6] Unlike the participants in the experiment, 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.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

faithful 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 the correct ranking of these SM constraints, 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).

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 (van Rossum 1995). 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 behaviour in a number of such tasks (Kirov 2017). For example, when tested on the experimental results from Albright & Hayes (2003), a Seq2Seq model's predictions correlated with human behaviour better 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 organised 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 the set of standard phonological features in (11). Since the model uses real-numbered values in its input and output, I used 1, −1 and 0 to represent +, − and unmarked feature values respectively.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

(11)

	[syllabic]	[voice]	[Coronal]	[anterior]	[Dorsal]	[high]
[t]	−	−	+	+		
[d]	−	+	+	+		
[ʧ]	−	−	+	−		
[ʤ]	−	+	+	−		
[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 items in the 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 four 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 was given a UR, such as /ti/, then its MSE was calculated for both the correct SR, [ʧi], and the incorrect choice that participants were presented with in the experiment, [ti]. If the model learned the correct UR → SR mapping, 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 (12).[7]

(12) *Accuracy estimation* (*Seq2Seq model*)

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

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

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

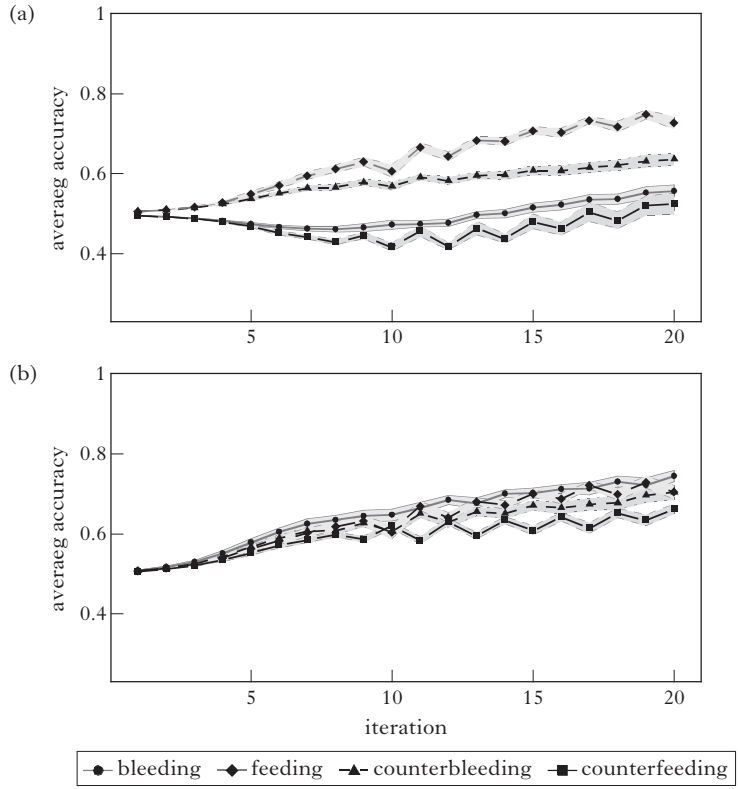


Figure 4

Average accuracy in (a) palatalising trials and (b) interacting trials for the Seq2Seq across iterations in a forced-choice task. Results are broken down by pattern type.

I used this as an estimate of the model's accuracy across each trial type in each language. Figure 4 shows 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 palatalising contexts (with the exception of the feeding language, which again took some time to become more accurate than its opaque counterparts on interacting items). And, like the EDL model, transparent languages tended to have higher accuracies in the interacting trials, while maximally utilising languages tended to be more accurate in palatalising trials. To confirm these qualitative results, I again performed two-way *t*-tests on the model's results from the end of learning (see Fig. 5 for a presentation of these results that is comparable to Fig. 1). These showed that both transparent languages ($t[78] = 2.452$, $p = 0.016$) and maximally utilising languages ($t[78] = 6.98$, $p < 0.0001$) had significantly higher accuracies in the relevant trial types.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

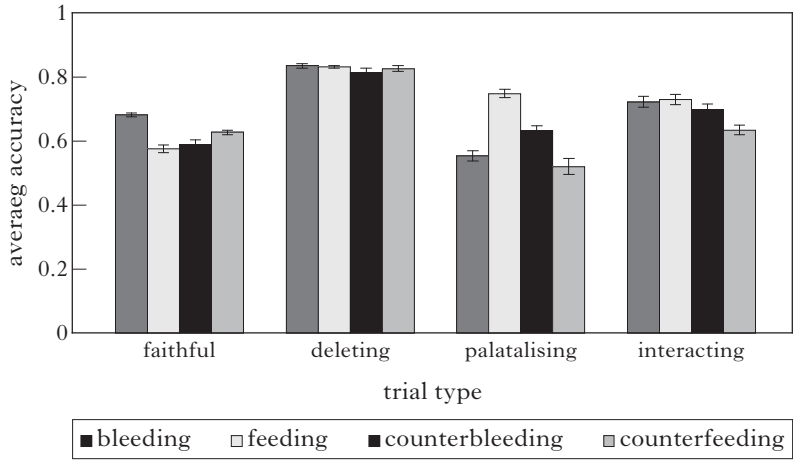


Figure 5

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 in 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 reasons 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 (see e.g. 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 is beyond 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 the biases proposed by Kiparsky (1968, 1971) emerge in the Seq2Seq model's learning, despite the lack of handmade 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 analysed the patterns they were learning in a number of other ways. For example, they could have learned the

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

differences across suffix and stem SRs 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 palatalisation rule applying to all /t/'s in the counterbleeding language). Because this study was concerned with testing whether the biases in Kiparsky (1968, 1971) occurred in human learning (regardless of the analysis that those speakers used), figuring out the exact pattern that participants learned is beyond the scope of this paper. However, future work could explore this topic more carefully, by testing participants' generalisation 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 URs 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 modelling the diachronic changes observed by Kiparsky (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 behaviour observed here. Nazarov & Pater's (2017) Stratal MaxEnt model, Rasin *et al*.'s (2017) minimum description length learner and general-purpose learning algorithms for finite state automata (e.g. 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, it did not affect the learning of the rules themselves. Instead, a MaxUtil bias was the predominant factor that affected how easily participants learned the palatalisation rule in the experiment. This localised 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 underutilised in training, the MaxUtil bias observed here could not have affected their participants' learning.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

The results in this paper also help to shed light on the inherent biases present in Jarosz' (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 show 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 proposed by Kiparsky (1971). 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 utilisation 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 to a model of phonological learning – challenging Kiparsky's (1971) assumption to the contrary.

REFERENCES

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

Baković, Eric (2011). Opacity and ordering. In John Goldsmith, Jason Riggle & Alan Yu (eds.) *The handbook of phonological theory*. 2nd edn. Malden, Mass. & Oxford: Wiley-Blackwell. 40–67.

Bates, Douglas, Martin Maechler, Ben Bolker & Steven Walker (2015). lme4: linear mixed-effects models using 'Eigen' and S4. R package (version 1.1-12). cran.r-project.org/web/packages/lme4.

Brooks, K. Michael, Bozena Pajak & Eric Baković (2013). Learning biases for phonological interactions. Poster presented at the 2013 Annual Meeting on Phonology, University of Massachusetts Amherst. Available (August 2019) at http://idiom.ucsd.edu/~bakovic/bpb/BPB-Phonology2013.pdf.

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

Chandlee, Jane & Adam Jardine (2014). Learning phonological mappings by learning Strictly Local functions. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.) *Proceedings of the 2013 Annual Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v1i1.13.

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

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

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes & Willem Zuidema (2018). Under the hood: using diagnostic classifiers to investigate and improve how language models track agreement information. *ArXiv Preprint*. http://arxiv.org/pdf/1808.08079.pdf.
Jarosz, Gaja (2014). Serial markedness reduction. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.) *Proceedings of the 2013 Annual Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v1i1.40.
Jarosz, Gaja (2015). Expectation driven learning of phonology. Ms, University of Massachusetts Amherst.
Jarosz, Gaja (2016a). Learning opaque and transparent interactions in Harmonic Serialism. In Gunnar Ólafur Hansson, Ashley Farris-Trimble, Kevin McMullin & Douglas Pulleyblank (eds.) *Proceedings of the 2015 Annual Meeting on Phonology*. http://dx.doi.org/10.3765/amp.v3i0.3671.
Jarosz, Gaja (2016b). Refining UG: connecting phonological theory and learning. Paper presented at the 47th Meeting of the North East Linguistic Society, University of Massachusetts Amherst. Slides available (August 2019) at https://blogs.umass.edu/jarosz/files/2016/10/NELS2016_final_red.pdf.
Kenstowicz, Michael & Charles Kisseberth (1971). Unmarked bleeding orders. *Studies in the Linguistic Sciences* **1:1**. 8–28.
Kim, Yun Jung (2012). Do learners prefer transparent rule ordering? An artificial language learning study. *CLS* **48:1**. 375–386.
Kiparsky, Paul (1968). Linguistic universals and linguistic change. In Emmon Bach & Robert T. Harms (eds.) *Universals in linguistic theory*. New York: Holt, Rinehart & Winston. 170–202.
Kiparsky, Paul (1971). Historical linguistics. In William Orr Dingwall (ed.) *A survey of linguistic science*. College Park: University of Maryland Linguistics Program. 576–642.
Kiparsky, Paul (2000). Opacity and cyclicity. *The Linguistic Review* **17**. 351–365.
Kirov, Christo (2017). Recurrent neural networks as a strong domain-general baseline for morpho-phonological learning. Poster presented at the 91st Annual Meeting of the Linguistic Society of America, Austin. Available (August 2019) at https://ckirov.github.io/papers/lsa2017.pdf.
Kirov, Christo & Ryan Cotterell (2018). Recurrent neural networks in linguistic theory: revisiting Pinker & Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics* **6**. 651–666.
Kurtz, Kenneth J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review* **14**. 560–576.
Li, Jiwei, Xinlei Chen, Eduard Hovy & Dan Jurafsky (2015). Visualizing and understanding neural models in NLP. *ArXiv Preprint*. http://arxiv.org/pdf/1506.01066.pdf.
Luce, R. Duncan (1959). *Individual choice behavior: a theoretical analysis*. New York: Wiley.
McCarthy, John J. (1999). Sympathy and phonological opacity. *Phonology* **16**. 331–399.
McCarthy, John J. (2007). *Hidden generalizations: phonological opacity in Optimality Theory*. Sheffield & Bristol, Conn.: Equinox.
McCoy, R. Thomas, Robert Frank & Tal Linzen (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *ArXiv Preprint*. http://arxiv.org/pdf/1802.09091.pdf.
Moreton, Elliott & Joe Pater (2012a). Structure and substance in artificial-phonology learning. Part 1: Structure. *Language and Linguistics Compass* **6**. 686–701.
Moreton, Elliott & Joe Pater (2012b). Structure and substance in artificial-phonology learning. Part 2: Substance. *Language and Linguistics Compass* **6**. 702–718.
Moreton, Elliott & Katya Pertsova (2016). Implicit and explicit processes in phonotactic learning. In Jennifer Scott & Deb Waughtal (eds.) *Proceedings of the 40th Annual*

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press

Boston University Conference on Language Development. Somerville, Mass.: Cascadilla. 277–290.
Nazarov, Aleksei & Joe Pater (2017). Learning opacity in Stratal Maximum Entropy Grammar. *Phonology* **34**. 299–324.
Pullum, Geoffrey K. (1976). The Duke of York gambit. *JL* **12**. 83–102.
Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted (2003). Phonological rule-learning and its implications for a theory of vowel harmony. *WCCFL* **22**. 423–435.
R Core Team (2016). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org.
Rahman, Fariz (2016). seq2seq: sequence to sequence learning with Keras. Available (August 2019) at https://github.com/farizrahman4u/seq2seq.
Rasin, Ezer, Iddo Berger, Nur Lan & Roni Katzir (2017). Rule-based learning of phonological optionality and opacity. Paper presented at the 48th Meeting of the North East Linguistic Society, University of Iceland. Available (August 2019) at http://www.mit.edu/~rasin/files/abstracts/nels2017.pdf.
Rossum, Guido van (1995). *Python tutorial*. Amsterdam: Centrum voor Wiskunde en Informatica.
Sanders, Nathan (2003). *Opacity and sound change in the Polish lexicon*. PhD dissertation, University of California, Santa Cruz.
Sutskever, Ilya, Oriol Vinyals & Quoc V. Le (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*. 3104–3112. Available (August 2019) at https://arxiv.org/abs/1409.3215v3.
Tieleman, Tijmen & Geoffrey Hinton (2012). Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* **4**. 26–31.
Wilson, Colin (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* **30**. 945–982.
Zuraw, Kie (2003). Probability in language change. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.) *Probabilistic linguistics*. Cambridge, Mass.: MIT Press. 139–176.

https://doi.org/10.1017/S0952675719000320 Published online by Cambridge University Press