Modelling a subregular bias in phonological learning with Recurrent Neural Networks

Brandon Prickett
University of Massachusetts Amherst

ABSTRACT

A number of experiments have demonstrated what seems to be a bias in human phonological learning for patterns that are simpler according to Formal Language Theory (Finley and Badecker 2008; Lai 2015; Avcu 2018). This paper demonstrates that a sequence-to-sequence neural network (Sutskever *et al.* 2014), which has no such restriction explicitly built into its architecture, can successfully capture this bias. These results suggest that a bias for patterns that are simpler according to Formal Language Theory may not need to be explicitly incorporated into models of phonological learning.

Keywords: neural networks, learning bias, Formal Language Theory, phonology

INTRODUCTION 1

Formal Language Theory (FLT; Chomsky 1956) describes how complex a pattern is in terms of the computational machinery needed to represent it. The framework was originally designed to demonstrate that natural language syntax was more complex than the set of *Regular* patterns (i.e., those that could be represented using finite state machines). However, Johnson (1972) showed that all known phonological mappings could be considered, at most, Regular (see also Kaplan and Kay

1994). Recent work has supported this finding, arguing that phonological learning must be categorically limited to patterns that can be characterized as *Subregular* (i.e., belonging to specific classes of patterns that can be represented with less expressive power than that of a finite state machine; Heinz 2010; Heinz and Idsardi 2011). One piece of evidence for this hypothesis is a series of experimental results that show humans being biased against learning certain patterns that seem to be too complex according to FLT-based metrics (Finley and Badecker 2008; Lai 2015; Finley 2017; Avcu 2018).

For example, Finley and Badecker (2008) showed that their participants were biased against learning *Majority Rule Harmony* (also known as *Majority Rules*; Lombardi 1999; Bakovic 2000), an unattested phonological process that is more complex than the set of Regular mappings. Later experimental work went on to show that people were also biased against learning some Subregular patterns (Lai 2015; Avcu 2018; McMullin and Hansson 2019), providing evidence that the phonological grammar might be limited to even simpler levels of the FLT hierarchy, such as those that can be characterized as *Strictly Local* and *Tier-based Strictly Local* (TSL; Heinz *et al.* 2011).[1] The former level of complexity includes any pattern that bans a finite set of substrings from occurring in a word, while the latter does so over a tier of segments (i.e., certain segments can be ignored by the pattern).

An example of a Strictly Local pattern that commonly occurs in natural language is the restriction banning voiceless sounds after nasals (henceforth *NC̥; Pater, 1999). This pattern is Strictly Local since it bans any word containing the finite set of strings that result from combining all nasals with all voiceless sounds (e.g. [nt], [np], [mt], [mp], etc.). TSL patterns are also common in phonology and are typically called *harmony* (see Rose and Walker 2011 for an overview), since many of them cause a subset of segments in a word to agree in their value for some feature.[2] For example, Navajo contains a har-

[1] *Strictly Piecewise* has also been suggested as an appropriate level of complexity to describe phonological patterns (Heinz 2010); however, see McMullin (2016) and Lamont (2018, 2019a) for arguments against this.

[2] Long-distance dissimilation patterns (i.e., patterns in which sounds must disagree in their value for a feature; Bennett 2015), are rarer in natural language but are also Tier-based Strictly Local.

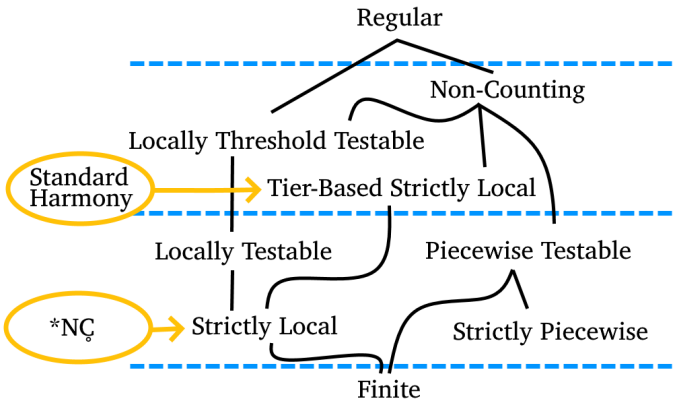

Figure 1:
The Subregular Hierarchy (Heinz 2018), with examples of Strictly Local and TSL patterns given. Dashed blue lines indicate different orders of logic. Solid black lines indicate subset relationships

mony pattern in which all sibilants (e.g., [s] and [ʃ]) within a word have to agree in their value for the feature [anterior] (Sapir and Hoijer 1967). This means that on the sibilant tier, the strings [sʃ] and [ʃs] are banned, since [s] is [+anterior] but [ʃ] is [−anterior]. Any sounds that are not sibilants are irrelevant to the pattern. A word like *[saʃ] would not be allowed, since its sibilant tier would exclude [a] and only include the banned sequence *[sʃ]. Figure 1 shows the full Subregular Hierarchy and where each of these two types of patterns are located in it.

While a considerable amount of work has been done to explain phonological typology and learning in terms of these FLT-based criteria, little work has been done to computationally model the experimental results that support a bias for Subregular patterns.[3] Here, I will show that the biases observed in past FLT-related experiments can emerge from the learning process of a relatively generic learner, namely a sequence-to-sequence neural network, which has the expressive power to represent both Subregular and Supraregular patterns (Siegelmann 1999). Since the network has no explicit, FLT-related biases built into its architecture, this provides evidence that such a

[3] Note that most of the literature involving FLT and learning (e.g., Chandlee *et al.* 2015; Jardine and Heinz 2016, among others) does not have an explicit hypothesis for how such learning algorithms can be used to make predictions for artificial language learning experiments. Instead, such work tends to focus on whether formally defined classes of languages are learnable at all, given certain kinds of training data.

bias may not need to be added to theories of phonological acquisition.

The paper is structured as follows: Section 2 introduces the neural network model that I will be using, Section 3 focuses on simulating experimental results regarding Majority Rule Harmony (Lombardi 1999; Bakovic 1999; Finley and Badecker 2008), Section 4 focuses on doing the same for experiments that involve First-Last Assimilation (Lai 2015; Avcu 2018), and Section 5 concludes.

2 MODELLING PHONOLOGICAL LEARNING WITH NEURAL NETWORKS

Neural networks have been used to model linguistic patterns since at least Rumelhart and McClelland (1986) and were quickly applied to the domain of phonology by Touretzky (1989) and Touretzky and Wheeler (1990). Hare (1990) first used *recurrent* neural networks (Jordan 1986; Elman 1990) to capture Hungarian vowel harmony, demonstrating that this architecture could be particularly useful for learning phonological mappings. Recurrent neural networks treat a stimulus as being made up of multiple timesteps, each of which the model processes separately. At each timestep, the model has connections that lead to the output layer and to the next step in time. These connections that feed into future timesteps are called recurrent and give the model a kind of memory as it walks through the full stimulus. This is illustrated in Figure 2 for Hungarian vowel harmony.

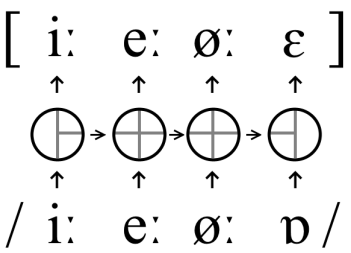

Figure 2:
Illustration of a recurrent neural network. Circles represent the hidden recurrent layer processing each timestep, black arrows represent groups of connections, grey lines represent the internal structure of the layer, and IPA symbols represent feature vectors corresponding to each segment

The use of such simple recurrent networks was later expanded to model other phonological phenomena, such as voicing assimilation (Gasser and Lee 1992) and phonotactic learning biases (Doucette 2017). However, these simple networks have been critiqued for their inability to generalise in a human-like way (Gasser 1993; Marcus *et al.* 1999) and for being too myopic (Alderete and Tupper 2018), since they have no ability to look ahead in their input sequence. There are a number of other reasons to suspect that simple recurrent networks would not be able to handle the full wealth of phonological phenomena – for example, their dependency on input and output lengths being equal (Sutskever *et al.* 2014).

Most of these issues are solved by the neural network architecture used in this paper, *sequence-to-sequence networks* (henceforth Seq2Seq; Sutskever *et al.* 2014). Seq2Seq networks were originally designed for machine translation and are meant to handle the fact that different languages often use different numbers of words to express the same idea. For example, a sentence like "No, I am your father" could be translated to Spanish as "No, soy tu padre," which has one less word. Seq2Seq networks deal with this by processing sequences in the input with a recurrent network called the *encoder* which is connected to a separate network, called the *decoder*, via its hidden layer connections. This processed data is then unpacked by the decoder into an output sequence whose length is independent of the length of the input.

This design also makes Seq2Seq networks well suited for modelling morphological and phonological patterns (e.g., Kirov and Cotterell 2018; Prickett *et al.* 2018; Prickett 2019), since these often involve mapping between forms of different lengths. For the simulations presented in this paper, words are represented as sequences of sounds, where sounds are vectors of real-numbered features that range from 0 to 1. In the input, which represents the underlying form, standard phonological features are used (like [voice] or [back]), with 0 and 1 corresponding to [−] and [+], respectively. In the output, which represents the surface representation, the network has a binary classifier for each feature that gives the model's estimated probability for how likely that feature is to have a positive value, given the underlying representation (UR) in its input. This is illustrated in Figure 3 using the same Hungarian example as above, with the feature vectors in the

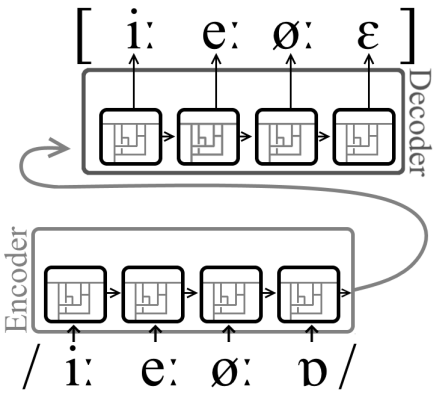

Figure 3:
An example of how Hungarian vowel harmony might be handled by a Seq2Seq network. The IPA symbols shown at the top and bottom of the figure represent the model's output and input, respectively, and stand in for vectors of real-numbered feature values. Black squares are Gated Recurrent Units and black arrows are sets of connections. The grey arrow shows the encoder's hidden layer activations being passed to the decoder

input and the most probable sets of feature values in the output being represented using IPA symbols.

The network presented here also uses *Gated Recurrent Units* (GRU; Cho *et al.* 2014) which were designed to solve another issue with simple recurrent networks: *vanishing gradients* (Bengio *et al.* 1994), which can prohibit a network from learning long-distance dependencies. While none of the patterns I investigate have dependencies that are long enough to be affected by this phenomenon, GRU units are relatively standard in the Seq2Seq literature and I leave it to future work to see whether they are necessary for capturing the results presented here. Similarly, in all of my simulations, the network's weights were optimized using Adam (Kingma and Ba 2015), a standard algorithm for training neural networks, but one that is likely not necessary to produce the results that I observed. The loss function used for optimization was the sum of binary cross entropy over all of the binary feature classifiers in the output and weight updates were made after seeing each word in training (i.e. batch sizes were equal to 1, sometimes called *online learning* in the phonological literature).

A final aspect of the model's architecture worth noting is *attention* (Bahdanau *et al.* 2015). This gives the model's decoder additional access to information from the input sequence by allowing it to see the decoder's hidden-state activations. Attention has been shown to encourage human-like generalization in Seq2Seq networks (Nelson *et al.* 2020). Some pilot simulations without attention suggested that it helped the model generalisese better in the simulations presented here.

MAJORITY RULE HARMONY 3

Background 3.1

Majority Rule Harmony is a pattern predicted by some constraint-based theories of assimilation in which the number of segments in a word's underlying representation (UR) with a particular feature value determines what the value of that feature will be throughout the surface representation (SR) of the word (Lombardi 1999; Bakovic 1999). For example, if a UR has two [−anterior] segments and only one [+anterior] segment (e.g. /saʃaʃ/), then the surface representation of the word would assimilate all of the sounds to be [−anterior] (e.g. [ʃaʃaʃ]). Conversely, if a UR has two [+anterior] sounds and only a single [−anterior] one (e.g. /sasaʃ/), the surface form would instead assimilate all of the sibilants to be [+anterior] (e.g. [sasas]). Since Majority Rule requires a potentially unbounded amount of memory (i.e. enough memory to keep track of the quantities for each feature value), it cannot be represented with a finite state transducer and is more complex than the set of Regular functions (Heinz and Lai 2013).[4]

Finley and Badecker (2008) tested whether humans were biased against Majority Rule. They did this by training participants on a language that was ambiguous between Majority Rule Harmony and a more standard, attested harmony pattern (henceforth *Attested Harmony*), in which the value of the relevant feature in the SR was determined by the value of that feature in either the leftmost or rightmost segment of the UR (see Rose and Walker 2011, for more on the kinds of harmony patterns that are common in natural language). Directional harmony mappings like this are Subregular, since determining how a vowel will surface only depends on local information in the input and

[4] Since TSL only defines a set of languages (i.e. phonotactic restrictions on SRs) and not a set of functions (i.e. UR→SR mappings), standard harmony patterns (when represented as transformations) are *Output Tier-based Strictly Local* (Burness and McMullin 2019), a subset of Regular *functions*. See Lamont (2019b) for more on this distinction between mappings and phonotactics and its relevance to complexity in phonology.

output (Chandlee 2014; Chandlee *et al.* 2014, 2015; Graf and Mayer 2018; Burness and McMullin 2019).

Participants in the experiment were exposed to stimuli meant to represent underlying forms like /kupoki/, with both [+back] and [−back] vowels present in a single word. Crucially, the minority vowel (/i/ in this case, since it is [−back] while /o/ and /u/ are both [+back]) always occurred on the same side of the word in training. After being given each "underlying" form, participants would then be exposed to a stimulus representing the "surface" form it mapped to (e.g., [kupoku] for the example above). The mapping /kupoki/→[kupoku] could then be analysed by the participants in two ways: either Attested Harmony, where the [back] value of the final vowel changed because the leftmost vowel in the word was [+back], or Majority Rule Harmony, where the word-final /i/ changed because the majority of vowels in the underlying form were [+back].

After being exposed to a number of these ambiguous mappings, participants were asked to choose between mappings that were unambiguous between Majority Rule and Attested Harmony.[5] For example, they might be given /kupeki/ and need to choose between mapping it to [kupoku] (the Attested Harmony candidate) or [kipeki] (the Majority Rule candidate). If participants chose between the options at chance, it would suggest that they had no preference for either pattern. However, if they chose one significantly more often than the other, it would suggest that they were biased toward learning that pattern. Finley and Badecker (2008) found that their participants were significantly more likely to generalise in a way that adhered to Attested Harmony. That is, when choosing to either apply an Attested Harmony or Majority Rule mapping to items that were unambiguous between the two patterns, participants only applied the latter in approximately 20% of trials. This suggests that in the face of ambiguous training, the participants learned the Attested Harmony pattern – which Finley and Badecker (2008) interpreted as

[5] Thanks to a reviewer for pointing out that these forms are only unambiguous as to which of the two patterns of interest they adhere to. A number of other analyses could be used to account for both sets of words, such as a bidirectional harmony process for the Majority Rule items (where the value of [back] spreads outward from the middle vowel).

evidence of a bias against learning Supraregular patterns like Majority Rule.

Simulations 3.2

To see whether the behaviour observed by Finley and Badecker (2008) is mirrored by a Seq2Seq network, I simulated their experiment using the architecture described in Section 2. The model was exposed to the same types of training data that Finley and Badecker (2008) gave their participants, which was ambiguous between Majority Rule and Attested Harmony. Since only the vowels were relevant to the patterns in this experiment, all consonants were removed. Other than this difference, the model was exposed to the same underlying and surface forms that the experiment participants were given. These are shown in Table 1 and the features used in all the simulations presented in this subsection are shown in Table 2.

All simulations consisted of 15 repetitions using this training data, with randomly initialized weights at the start of learning, and 300 full passes through the training data (i.e., 300 epochs). At each epoch, the

Underlying Representation	Surface Representation
/o u i/	[o u u]
/e i o/	[e i e]
/u o i/	[u o u]
/i e o/	[i e e]
/o u e/	[o u o]
/u o e/	[u o o]
/e i u/	[e i i]
/i e u/	[i e i]

Table 1:
Training data for Majority Rule simulations

	[back]	[high]
i	−	+
u	+	+
e	−	−
o	+	−

Table 2:
Features for Majority Rule simulations

Table 3: Test Data for Majority Rule simulations. Model was given a UR as input (shown in the leftmost column) and assigned probabilities to each output choice (shown in the center and rightmost columns)

UR	Attested Harmony SR	Majority Rule SR
/o i e/	[o u o]	[e i e]
/o e i/	[o o u]	[e e i]
/u i e/	[u u o]	[i i e]
/u e i/	[u o u]	[i e i]
/i o u/	[i e i]	[u o u]
/i u o/	[i i e]	[u u o]
/e o u/	[e e i]	[o o u]
/e u o/	[e i e]	[o u o]

model was presented with the same kind of crucial forced choices that Finley and Badecker (2008) gave their participants in the experiment's test phase (shown in Table 3).

The conditional probability that the model assigned to each choice, given a particular UR, was calculated using the equation defined in Equation 2, based on Luce (1959), where $pr(UR_i) \rightarrow SR_j$ is found using Equation 1, and where f_{ij} stands for feature j in segment s_i of the relevant SR.

$$(1) \qquad pr(UR \rightarrow SR) = \prod\prod\prod pr(f_{ij}|UR)$$

$$(2) \quad pr(UR_i \rightarrow SR_1 | SR_1 \text{ or } SR_2) = \frac{pr(UR_i \rightarrow SR_1)}{pr(UR_i \rightarrow SR_1) + pr(UR_i \rightarrow SR_2)}$$

Results for these forced choice estimates were averaged over stimulus types and repetitions, and these averages are shown for each epoch in Figure 4. Figure 5 gives the 50th epoch in more detail, for results that are more visually comparable to the ones presented by Finley and Badecker (2008).

These results show that throughout learning, the model prefers choices that are consistent with Attested Harmony, even though it has been trained on data that is ambiguous between the two patterns. This difference reaches statistical significance for a range of epochs (including the 50th epoch), meaning that the bias in humans observed by Finley and Badecker (2008) can be captured by the model.

To further test the model's biases in regards to Majority Rule Harmony, I also ran a simulation that does not correspond to Finley and

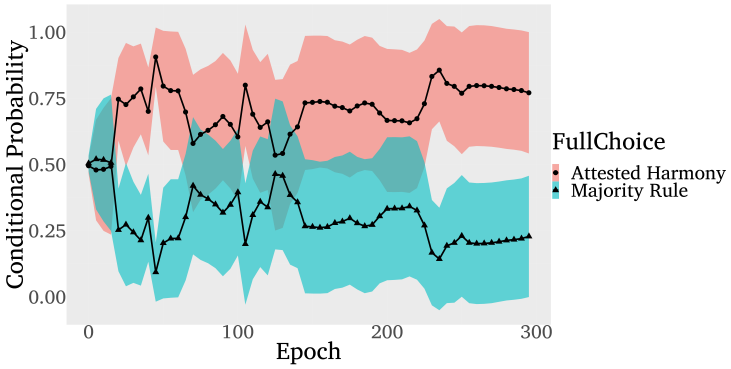

Figure 4:
Forced choice probabilities at each epoch in learning for the simulations of Finley and Badecker (2008). Coloured regions show 95% confidence intervals

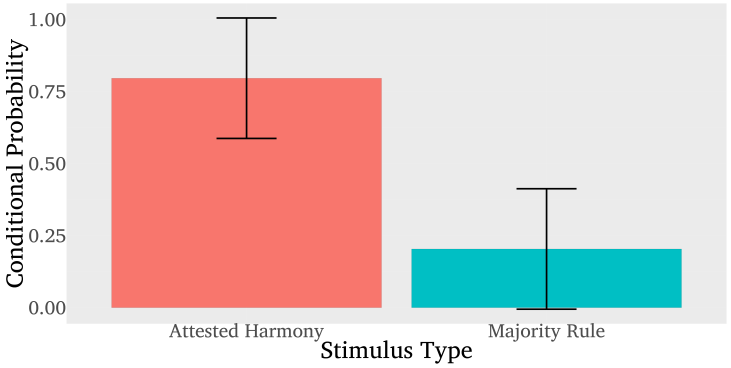

Figure 5:
Forced choice probabilities for the 50th epoch of training in the simulation of Finley and Badecker (2008). Error bars show 95% confidence intervals

Badecker's (2008) experiment. Rather than using a generalization-based design, in this simulation, multiple, unambiguous languages were used in training. Additional data points were added to the training data in Table 1 to disambiguate the two patterns of interest. The data for unambiguous versions of Majority Rule Harmony and Attested Harmony are shown in Tables 4 and 5.

The model was trained on these unambiguous versions of Attested Harmony and Majority Rule and the cross entropy and accuracy were recorded at each epoch. Accuracy was estimated by feeding the model each of the underlying forms in the training data, sampling from the probabilities it produced in the output to create surface forms, and finding the proportion of those surface forms that were perfectly produced in that epoch's sample. The learning curves created from these results (averaged over 15 repetitions) are shown in Figure 6.

These results show that for small portions of the learning curve, Attested Harmony's average accuracy is marginally higher than

Table 4:
Training data for the unambiguous Majority Rule language, based on the ambiguous data from Finley and Badecker (2008). Bolded cells show which data are unambiguous

Underlying Representation	Surface Representation
/o u i/	[o u u]
/e i o/	[e i e]
/u o i/	[u o u]
/i e o/	[i e e]
/o u e/	[o u o]
/u o e/	[u o o]
/e i u/	[e i i]
/i e u/	[i e i]
/o i e/	**[e i e]**
/o e i/	**[e e i]**
/u i e/	**[i i e]**
/u e i/	**[i e i]**
/i o u/	**[u o u]**
/i u o/	**[u u o]**
/e o u/	**[o o u]**
/e u o/	**[o u o]**

Table 5:
Training data for the unambiguous Attested Harmony language, based on the ambiguous data from Finley and Badecker (2008). Bolded cells show which data are unambiguous

Underlying Representation	Surface Representation
/o u i/	[o u u]
/e i o/	[e i e]
/u o i/	[u o u]
/i e o/	[i e e]
/o u e/	[o u o]
/u o e/	[u o o]
/e i u/	[e i i]
/i e u/	[i e i]
/o i e/	**[o u o]**
/o e i/	**[o o u]**
/u i e/	**[u u o]**
/u e i/	**[u o u]**
/i o u/	**[i e i]**
/i u o/	**[i i e]**
/e o u/	**[e e i]**
/e u o/	**[e i e]**

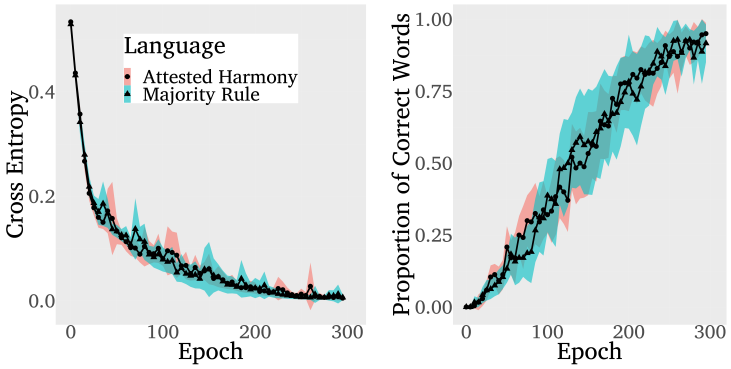

Figure 6: Learning curves for Majority Rule and Attested Harmony in the simulations using unambiguous versions of the language from Finley and Badecker (2008). Chance performance for the plot on the right would be considerably lower than 0.1, since the model assigns probabilities to each feature value in each segment. Coloured regions show 95% confidence intervals

Majority Rule's, but this difference is not a reliable one. There also seems to be a small, statistically marginal difference between the loss curves for the two patterns, but this effect is even less consistent throughout learning. Assuming that the small, artificial languages used here adequately represented each of the languages, this suggests that if the model does have a bias for Subregular patterns in its learning from unambiguous data, the effect size of this bias is too small to see in just 15 repetitions.

FIRST-LAST ASSIMILATION 4

Background 4.1

First-Last Assimilation is a hypothetical phonotactic restriction in which the first and last segment of a word must agree in some feature value, while the intervening sounds are ignored (Lai 2015). For example, if the feature that needed to agree was [anterior], the word [saʃas] would be allowed, but the word *[saʃaʃ] would not be. Lai (2015) argued that there are reasonable diachronic origins for such a pattern,

since the beginning and end of a word are perceptually salient positions. She went on to argue that the absence of such a pattern in the phonological typology could be due to its FLT-based complexity.

While First-Last Assimilation is Subregular, it belongs to the *Locally Testable* region, which is more complex than TSL, in terms of the logic needed to define the crucial parts of the pattern. That is, sets of sequences are necessary to describe words banned by First-Last Assimilation (i.e. "words with either [#s] *and* [ʃ#] or [#ʃ] *and* [s#] are banned"), which is never true for TSL patterns.

Two studies have shown that people have biases against learning First-Last Assimilation. Lai (2015) trained participants on either a standard sibilant harmony pattern (henceforth, *Attested Harmony*) or First-Last Assimilation by having them listen to and then repeat words adhering to the pattern they were assigned to. In the testing phase of the experiment, participants were asked to judge which word was more likely to belong to the language they were trained on in three types of forced choice:[6]

i. a choice between a word that was allowed in both patterns (e.g. [sasakas], denoted as FL/AH below) and a word that was only allowed in First-Last Assimilation (e.g. [saʃakas], denoted as FL/*AH below),
ii. a choice between a word that was allowed in both patterns and a word that was banned by both (e.g. [sasakaʃ], denoted as *FL/*AH below),
iii. a choice between a word that was only allowed in First-Last Assimilation and one that was banned by both.

Participants who learned an Attested Harmony pattern would be expected to choose words that were allowed by both patterns when presented with choices (i) and (ii), but should choose at random for choice (iii). This is because choice (iii) forces participants to choose between two words that are both banned by the Attested Harmony pattern. Participants who learned a First-Last Assimilation pattern would

[6] While there are more than three logically possible forced choice options, including words that were only allowed in Attested Harmony would have been impossible. This is because all words that are allowed in Attested Harmony are also allowed in First-Last Assimilation.

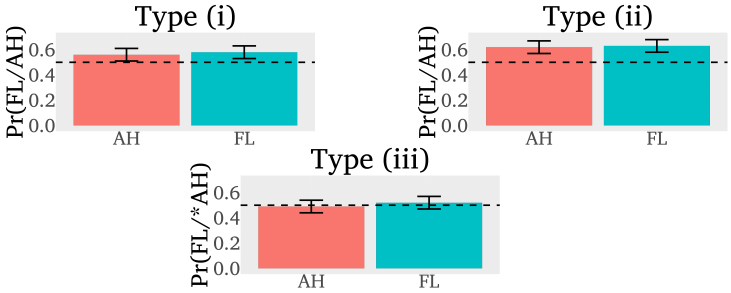

Figure 7: Results adapted from Figures 2–4 in Lai (2015). The x-axis shows which pattern participants were trained on. Type labels are mine, with "FL" standing for First-Last Assimilation, "AH" for Attested Harmony, and "*" indicating an option not being allowed in a given pattern. Note that Lai (2015) used the term "Standard Harmony"/"SH" for the pattern I'm calling "Attested Harmony"/"AH"

be expected to choose at chance for choice (i), since both choices are grammatical according to First-Last Assimilation. For choice (ii), they would be expected to choose words that are allowed by both patterns, and for choice (iii) they should choose the words that are only allowed by First-Last Assimilation.

However, participants trained on First-Last Assimilation in Lai's (2015) experiment did not behave as expected. Her results (reproduced in Figure 7) showed that participants in both language conditions behaved as if they had learned Attested Harmony.

Specifically, when presented with choices (i) and (ii), participants in both conditions chose items that were grammatical in both languages significantly more than chance, showing that they preferred items in which Attested Harmony was not violated. However, when presented with choice (iii), participants performed at chance, demonstrating that they had no preference between items that violated First-Last Assimilation and those that did not. This shows that they failed to learn First-Last Assimilation when trained on the pattern, and instead learned the Attested Harmony pattern. These results are what one would expect if there were a categorical restriction banning the acquisition of phonological patterns that are more complex than TSL.

Avcu (2018) ran another artificial language learning experiment to test for a bias against First-Last Assimilation. Participants received the same training as Lai's (2015) study; however in testing, they were asked to make a different kind of choice. Instead of choosing between

two words, participants judged whether they thought each test stimulus (some of which followed the pattern from training and some which did not) belonged to the language they had just learned. This allowed Avcu (2018) to analyse participant responses using *Signal Detection Theory* (Green and Swets 1966) and provided a measure of how sensitive individuals were to whether a word belonged to the language they were assigned. The results showed that participants in both language conditions were better than chance at performing this discrimination task, but that those who learned Attested Harmony performed significantly better. Since Avcu's (2018) participants were less successful at learning First-Last Assimilation than its more standard counterpart, these results also support the idea of a bias for patterns that are simpler according to FLT.

4.2 *Simulations*

To see if an explicit, FLT-related bias is needed to capture the results that Lai (2015) and Avcu (2018) observed in human learning, I ran a simulation using a Seq2Seq network.[7] The training and testing data that the model received were identical to the stimuli used by Lai (2015), except that all vowels were removed from the model's representations (as they were irrelevant to the patterns of interest).

Since Lai's (2015) participants were not exposed to the underlying forms for any of the stimuli, all training and testing data for the model assumed that underlying forms were identical to their corresponding surface forms (see Prince and Tesar 2004, for a similar approach to phonotactic learning). While this data represents an identity mapping, the fact that neural networks cannot perfectly learn such a mapping (Tupper and Shahriari 2016) means that the model must learn alternative ways to optimize its objective function, such as acquiring the phonotactic patterns present in the language (see Kurtz 2007, for a similar approach using a different neural network architecture). The

[7] Thanks to a reviewer for pointing me toward similar work in the domain of syntax: Ravfogel *et al.* (2019) show that a neural network, when trained on data that is ambiguous between an agreement pattern analogous to First-Last Assimilation and a pattern that involves more local agreement, the network generalises in a way that suggests it learned the latter.

Surface Representation
[ʃ s k ʃ]
[s ʃ k s]
[ʃ k s ʃ]
[s k ʃ s]
[ʃ ʃ k ʃ]
[s s k s]
[ʃ k ʃ ʃ]
[s k s s]

Table 6:
Training data for First-Last Assimilation language in the simulations of Lai (2015). The input and output to the model was identical for all data

Surface Representation
[ʃ ʃ k ʃ]
[s s k s]
[ʃ k ʃ ʃ]
[s k s s]
[ʃ ʃ k ʃ]
[s s k s]
[ʃ k ʃ ʃ]
[s k s s]

Table 7:
Training data for Attested Harmony language in the simulations of Lai (2015). The input and output to the model was identical for all data

	[anterior]	[sibilant]
s	+	+
ʃ	−	+
k	−	−

Table 8:
Features and segments used in Lai (2015) simulations

training data for First-Last Assimilation and Attested Harmony are shown in Tables 6 and 7, respectively. Additionally, the features used to represent the segments in both patterns are shown in Table 8.

Simulations consisted of 15 repetitions in each language condition, with randomly initialized weights at the start of learning, and 300 passes through the full data set. At each epoch of training, the model's cross entropy and accuracy were measured. Accuracy was estimated by feeding the model each of the forms in the training data as input, sampling from the probabilities it produced in its output to create surface forms, and finding the proportion of those surface forms

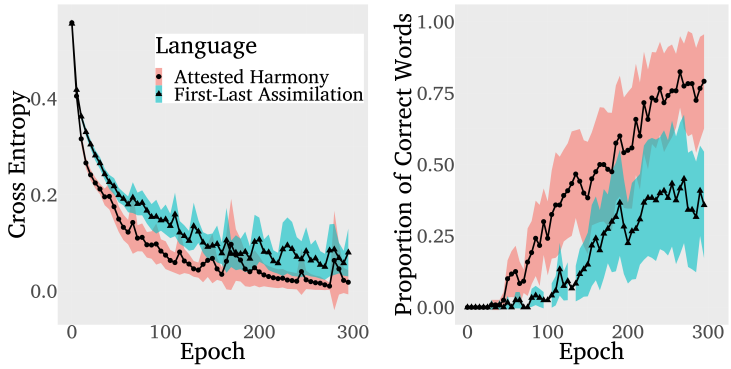

Figure 8: Learning curves for First-Last Assimilation and Attested Harmony in the simulations of Lai (2015). Chance performance for the plot on the right would be considerably lower than 0.1, since the model assigns probabilities to each feature value in each segment. Coloured regions show 95% confidence intervals

that matched their input in that epoch's sample. Learning curves showing both of these metrics are given in Figure 8.

The curves in Figure 8 show that Attested Harmony is learned consistently faster than First-Last Assimilation. This difference is significant for considerable portions of learning in both the model's loss and accuracy. These results are most comparable to those reported by Avcu (2018), since the model's performance is higher than chance for both patterns, but significantly better for Attested Harmony.

To compare the model's learning to the results in Lai (2015), the network was given a forced-choice task similar to the one described in Section 3.2, with the test data given in Table 9.

Since the patterns here were phonotactic (rather than mappings), there was no shared UR between the two choices. That is, the conditional probability that the model assigned to each choice was just a normalized probability for each of the two SRs mapping to themselves, as shown in Equation 3.

$$(3) \qquad pr(SR_1 | SR_1 \text{ or } SR_2) = \frac{pr(SR_1 \rightarrow SR_1)}{pr(SR_1 \rightarrow SR_1) + pr(SR_2 \rightarrow SR_2)}$$

The relevant conditional probabilities were averaged over stimulus types and repetitions, and are shown in Figure 9 and Figure 10 for the model that was trained on First-Last Assimilation and the model that was trained on Attested Harmony, respectively.

FL/*AH Choice	*FL/*AH Choice
[s k ʃ s]	[ʃ k ʃ s]
[ʃ s k ʃ]	[ʃ s k s]
[s ʃ k s]	[s ʃ k ʃ]
[ʃ k s ʃ]	[s k s ʃ]
FL/AH Choice	***FL/*AH Choice**
[s k s s]	[s k s ʃ]
[ʃ ʃ k ʃ]	[s ʃ k ʃ]
[ʃ k ʃ ʃ]	[ʃ k ʃ s]
[s s k s]	[ʃ s k s]
FL/AH Choice	**FL/*AH Choice**
[ʃ ʃ k ʃ]	[ʃ s k ʃ]
[s k s s]	[s k ʃ s]
[s s k s]	[s ʃ k s]
[ʃ k ʃ ʃ]	[ʃ k s ʃ]

Table 9:
Test data for First-Last Assimilation simulations. Probabilities for each form in the left column were normalized with their corresponding item in the right column. These normalized probabilities were then used to simulate the model's performance on the forced-choice task from Lai (2015)

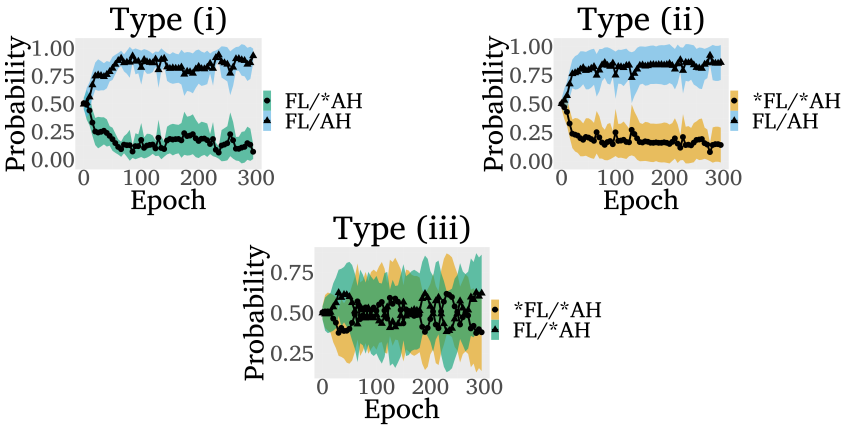

Figure 9: Forced choice probabilities at each epoch in learning for the First-Last Assimilation language. Coloured regions show 95% confidence intervals

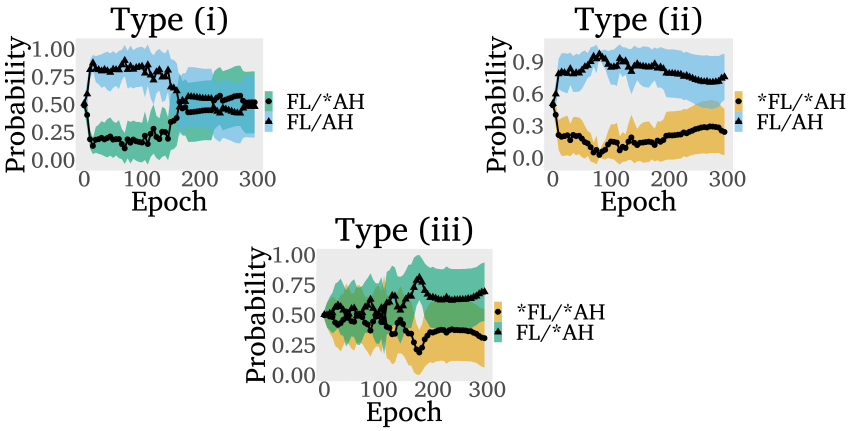

Figure 10: Forced choice probabilities at each epoch in learning for the Attested Harmony Language. Coloured regions show 95% confidence intervals

These results show that the Seq2Seq model, like the human participants in Lai (2015), behaved in a way that was consistent with Attested Harmony, even when trained on data that unambiguously followed the First-Last Assimilation pattern. That is, regardless of the model's training data, it chose at chance between words that were banned by Attested Harmony, even when one of those words adhered to First-Last Assimilation (with the only exception to this behaviour being a small number of epochs in the Attested Harmony condition). This is shown in the results for choice (iii). By itself, this only shows that the model did not learn First-Last Assimilation. However, choices (i) and (ii) both show that the models acquired Attested Harmony, since words adhering to this pattern are consistently given more probability than words banned by it for most of the acquisition process.[8] To show these results in a way that is more visually comparable to the results reported in Lai (2015), the model's estimates for the 100th

[8] Although, note that toward the end of learning, the model trained on the attested pattern begins to choose at chance in all three of the choice types. This could be due to the model eventually learning to faithfully map the segments in the input in those cases. While this approximates an identity mapping for the segments that were present in the training, it would not be a true identity mapping, since neural networks trained with algorithms like Adam cannot capture identity-based functions (Tupper and Shahriari 2016).

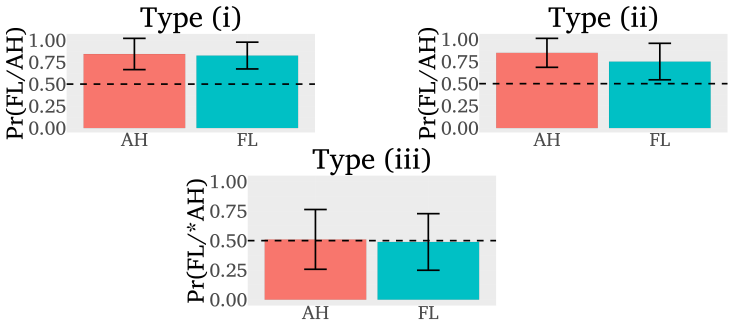

Figure 11: Forced choice probabilities for the 100th epoch of training in both the First-Last Assimilation language and the Attested Harmony Language. The dashed line shows chance and the error bars show 95% confidence intervals. As in Figure 7, "FL" stands for First-Last Assimilation, "AH" stands for Attested Harmony, and "*" indicates an option not being allowed in a given pattern

epoch in each language, which was a relatively representative point in each language's learning curve, are shown in Figure 11.

5 DISCUSSION

5.1 *Why can the Seq2Seq network capture these biases?*

In this paper, I showed that the apparent FLT-related bias observed in past artificial language learning experiments could be modeled by a recurrent neural network with no FLT-based restrictions built into its architecture. But the question of why these biases exist has not been addressed. One reason for the model's bias against Majority Rule Harmony could be its inability to count. Weiss *et al.* (2018) showed that GRU units, like the one used in the hidden layer of the neural network I tested, prohibit a model from acquiring the ability to count (as opposed to simple recurrent networks and networks with LSTM units, which were able to learn counting-based patterns). Since Majority Rule Harmony requires counting the occurrences of a particular feature value in the input, this could explain the model's preference

for learning an Attested Harmony pattern in the face of ambiguous data.

Another relevant factor is the locality bias (sometimes also called "sequentiality"; Battaglia *et al.* 2018) present in all recurrent network architectures. This is a bias for patterns that involve local dependencies, originating from the fact that recurrent connections have a finite amount of memory with which to store information across time. Past results on syntactic patterns have shown that this bias can cause RNNs to learn a local agreement pattern when given ambiguous evidence between that and a non-local one (Ravfogel *et al.* 2019). Similarly, McCoy *et al.* (2020) showed that Seq2Seq neural networks similar to the one used here were more likely to learn syntactic patterns that depended on linear order, which typically involves more local dependencies, than patterns that depended on hierarchical structure, which typically involves longer distance dependencies. Since First-Last Assimilation also involves non-local dependencies (i.e. two arbitrarily distant first and last segments), the network could have struggled to keep track of the relevant feature values in its recurrent connections when acquiring that pattern.[9]

5.2 *Future work*

This paper has shown that three experiments that found evidence supporting an FLT-based bias in humans (Finley and Badecker 2008; Lai 2015; Avcu 2018) can be simulated using a Seq2Seq recurrent neural network. Future work should continue to explore the phonological learning biases present in both humans and computational models. For example, one phonological pattern that was not discussed here but which the literature has discussed in detail is *Sour Grapes Harmony* (Bakovic 2000; Wilson 2003). Sour Grapes is identical to Standard Harmony, except when a segment that blocks the harmony process is

[9] The difference between local and non-local dependencies has been thoroughly explored in the statistical learning literature as well (e.g., Newport and Aslin 2004), and simulations of such statistical learning experiments with RNNs have been performed (see, e.g., Farkaš 2008). I leave exploring the relationship between these experiments and those that have been used to support FLT-based biases in phonology to future work.

present in a word. When this happens, any changes that would have occurred up to the blocker are prevented from occurring at all. Like First-Last Assimilation and Majority Rule, Sour Grapes is unattested in natural language and more complex than the Tier-based Strictly Local region of the Subregular Hierarchy (O'Hara and Smith 2019; Lamont 2019b).

Another avenue for future work is using more realistic artificial languages. In all of the experiments simulated here, word length was kept constant. When testing the effects of formal complexity on human learning, generalization to novel lengths has been shown to be crucial in understanding human bias (Westphal-Fitch *et al.* 2018). Further research that makes use of variable lengths in its training and testing data could shed light on whether humans display an FLT-based bias under these more realistic conditions.

Researchers should also explore how the predictions about human learnability made by FLT and neural networks differ. For example, certain Context-Sensitive patterns are easier for neural networks and humans to learn than corresponding Context-Free patterns (Li *et al.* 2013; Westphal-Fitch *et al.* 2018), despite the fact that Context-Sensitive is more complex according to FLT. Exploring whether mismatches like this occur in phonological patterns could shed more light on how psychologically real FLT-based complexity is.

Understanding better *why* the neural network is able to capture these results and what representations it learns while doing so is another important next step. While the interpretability of recurrent networks has primarily been explored in the context of syntactic patterns and language modelling (see, e.g., Alishahi *et al.* 2019, for a review), some recent work on phonological patterns has shown promising results in this direction (Nelson *et al.* 2020; Smith *et al.* 2021) and these techniques could likely be applied to the networks used here.

Finally, a number of choices about the model I used were made somewhat arbitrarily: the number of hidden states in each layer, the use of GRU instead of a different kind of recurrent layer in the model, the use of attention, *et cetera*. Changing any one of these would likely have an effect on the model's ability to capture the experiment results investigated in Section 3 and Section 4, and I leave exploring the consequences of such changes to future work.

5.3 *The relationship between FLT and other complexity metrics*

The Subregular Hierarchy is not the only way of measuring complexity that has been used in phonological research. Feature counting (Chomsky and Halle 1968), Minimum Description Length (Rasin and Katzir 2016), and various other methods (e.g. Moreton *et al.* 2017) have been used to characterize the complexity of phonological patterns. While these other methods are related to FLT, they are not perfectly correlated with it. For example, a feature-counting complexity metric would find a pattern banning all voiced sounds at the end of words (i.e., *[+voice]#) to be simpler than a pattern banning voiced, velar stops in that context (i.e., *[+voice, Dorsal]#). However, according to FLT, these patterns would both be Strictly Local, with no difference in complexity. Exploring the relationship between FLT and these other metrics is outside the scope of this paper; however future work should investigate what formalizations of complexity best predict both human behavior and linguistic typology (see, e.g., Moreton and Pater 2012).

5.4 *Conclusions*

Past work has explained phonological typology using an explicit, categorical restriction that prohibits the acquisition of patterns that are too complex according to the Subregular Hierarchy. Evidence for this hypothesis includes a series of experiments that showed humans being affected by an apparent FLT-based bias in an artificial language learning context (Finley and Badecker 2008; Lai 2015; Avcu 2018).

The results in this paper challenge the idea that a categorical, explicit bias like this is needed to capture phonological learning, since a Seq2Seq neural network with the expressive power to represent Supraregular patterns was able to capture these experimental results. While FLT can be useful for describing phonological typology, these results suggest that an explicit FLT-based bias may not be needed in models of phonological learning.

ACKNOWLEDGEMENT

The author would like to thank Joe Pater, Gaja Jarosz, John Kingston, Mohit Iyyer, Katya Pertsova, and Andrew Lamont for helpful discussion. Thanks also to the reviewers for their valuable feedback.

REFERENCES

John Alderete and Paul Tupper (2018), Connectionist Approaches to Generative Phonology, *The Routledge Handbook of Phonological Theory. Routledge*.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen (2019), Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop, *arXiv preprint arXiv:1904.04063*.

Enes Avcu (2018), Experimental Investigation of the Subregular Hypothesis, in *Proceedings of the 35th West Coast Conference on Formal Linguistics*, pp. 77–86.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2015), Neural Machine Translation by Jointly Learning to Align and Translate, in *3rd International Conference on Learning Representations, Conference Track Proceedings*.

Eric Bakovic (1999), Assimilation to the Unmarked, *University of Pennsylvania Working Papers in Linguistics*, 6(1):2.

Eric Bakovic (2000), *Harmony, dominance and control*, PhD Thesis, Rutgers University.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, *et al.* (2018), Relational Inductive Biases, Deep Learning, and Graph Networks, *arXiv preprint arXiv:1806.01261*.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi (1994), Learning Long-term Dependencies with Gradient Descent is Difficult, *IEEE Transactions on Neural Networks*, 5(2):157–166.

Wm G Bennett (2015), *The phonology of Consonants: Harmony, Dissimilation and Correspondence*, Cambridge University Press.

Phillip Burness and Kevin McMullin (2019), Efficient Learning of Output Tier-based Strictly 2-local Functions, in *Proceedings of the 16th Meeting on the Mathematics of Language*, pp. 78–90.

Jane Chandlee (2014), *Strictly Local Phonological Processes*, PhD Thesis, University of Delaware.

Jane Chandlee, Rémi Eyraud, and Jeffrey Heinz (2015), Output Strictly Local Functions, in *14th Meeting on the Mathematics of Language*, pp. 112–125.

Jane Chandlee, Rémi Eyraud, and Jeffrey Heinz (2014), Learning Strictly Local Subsequential Functions, *Transactions of the Association for Computational Linguistics*, 2:491–504.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (2014), On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Association for Computational Linguistics.

Noam Chomsky (1956), Three Models for the Description of Language, *IRE Transactions on Information Theory*, 2(3):113–124.

Noam Chomsky and Morris Halle (1968), *The Sound Pattern of English*, Harper & Row.

Amanda Doucette (2017), Inherent Biases of Recurrent Neural Networks for Phonological Assimilation and Dissimilation, in *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics*, pp. 35–40.

Jeffrey L. Elman (1990), Finding Structure in Time, *Cognitive science*, 14(2):179–211.

Igor Farkaš (2008), Learning Nonadjacent Dependencies with a Recurrent Neural Network, in *International Conference on Neural Information Processing*, pp. 292–299, Springer.

Sara Finley (2017), Locality and Harmony: Perspectives from Artificial Grammar Learning, *Language and Linguistics Compass*, 11(1):1–16.

Sara Finley and William Badecker (2008), Analytic biases for vowel harmony languages, in *West Coast Conference on Formal Linguistics*, volume 27, pp. 168–176.

Michael Gasser (1993), *Learning Words in Time: Towards a Modular Connectionist Account of the Acquisition of Receptive Morphology*, Indiana University, Department of Computer Science.

Michael Gasser and Chan-Do Lee (1992), Networks that Learn about Phonological Feature Persistence, in *Connectionist Natural Language Processing*, pp. 349–362, Springer.

Thomas Graf and Connor Mayer (2018), Sanskrit n-Retroflexion is Input-Output Tier-Based Strictly Local, in *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 151–160.

David Marvin Green and John A. Swets (1966), *Signal Detection Theory and Psychophysics*, volume 1, Wiley.

Mary Hare (1990), The Role of Trigger-target Similarity in the Vowel Harmony Process, in *Annual Meeting of the Berkeley Linguistics Society*, volume 16, pp. 140–152.

Jeffrey Heinz (2010), Learning Long-distance Phonotactics, *Linguistic Inquiry*, 41(4):623–661.

Jeffrey Heinz (2018), The computational nature of phonological generalizations, in *Phonological typology*, pp. 126–195, De Gruyter Mouton.

Jeffrey Heinz and William Idsardi (2011), Sentence and Word Complexity, *Science*, 333(6040):295–297.

Jeffrey Heinz and Regine Lai (2013), Vowel Harmony and Subsequentiality, in *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, pp. 52–63.

Jeffrey Heinz, Chetan Rawal, and Herbert G Tanner (2011), Tier-based Strictly Local Constraints for Phonology, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 58–64, Association for Computational Linguistics.

Adam Jardine and Jeffrey Heinz (2016), Learning Tier-based Strictly 2-local Languages, *Transactions of the Association for Computational Linguistics*, 4:87–98.

C. Douglas Johnson (1972), *Formal Aspects of Phonological Description*, Mouton & Co. N.V.

Michael I. Jordan (1986), Serial Order: A Parallel Distributed Processing Approach., Technical report, University of California, San Diego.

Ronald M. Kaplan and Martin Kay (1994), Regular Models of Phonological Rule Systems, *Computational Linguistics*, 20(3):331–378.

Diederik P. Kingma and Jimmy Ba (2015), Adam: A Method for Stochastic Optimization, in *3rd International Conference on Learning Representations, Conference Track Proceedings*.

Christo Kirov and Ryan Cotterell (2018), Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate, *Transactions of the Association for Computational Linguistics*, 6:651–665.

Kenneth J. Kurtz (2007), The Divergent Autoencoder (DIVA) Model of Category Learning, *Psychonomic Bulletin & Review*, 14(4):560–576.

Regine LAI (2015), Learnable vs. Unlearnable Harmony Patterns, *Linguistic Inquiry*, 46(3):425–451.

Andrew LAMONT (2018), Precedence is Pathological: The Problem of Alphabetical Sorting, *Proceedings of the 36th West Coast Conference on Formal Linguistics*, pp. 243–249.

Andrew LAMONT (2019a), Majority Rule in Harmonic Serialism, in *Proceedings of the Annual Meetings on Phonology*, volume 7.

Andrew LAMONT (2019b), Sour Grapes is Phonotactically Complex, Linguistic Society of America, 2019 Annual Meeting.

Feifei LI, Shan JIANG, Xiuyan GUO, Zhiliang YANG, and Zoltan DIENES (2013), The Nature of the Memory Buffer in Implicit Learning: Learning Chinese Tonal Symmetries, *Consciousness and cognition*, 22(3):920–930.

Linda LOMBARDI (1999), Positional Faithfulness and Voicing Assimilation in Optimality Theory, *Natural Language & Linguistic Theory*, 17(2):267–302.

R. Duncan LUCE (1959), *Individual Choice Behavior*, Dover Publications.

Gary MARCUS, Sugumaran VIJAYAN, S. Bandi RAO, and Peter M. VISHTON (1999), Rule Learning by Seven-month-old Infants, *Science*, 283(5398):77–80.

R. Thomas MCCOY, Robert FRANK, and Tal LINZEN (2020), Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-sequence Networks, *Transactions of the Association for Computational Linguistics*, 8:125–140.

Kevin MCMULLIN and Gunnar Ólafur HANSSON (2019), Inductive Learning of Locality Relations in Segmental Phonology, *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 10(1).

Kevin James MCMULLIN (2016), *Tier-based Locality in Long-distance Phonotactics: Learnability and Typology*, Ph.D. thesis, University of British Columbia.

Elliott MORETON and Joe PATER (2012), Structure and Substance in Artificial-phonology Learning, Part I: Structure, *Language and Linguistics Compass*, 6(11):686–701.

Elliott MORETON, Joe PATER, and Katya PERTSOVA (2017), Phonological Concept Learning, *Cognitive science*, 41(1):4–69.

Max NELSON, Hossep DOLATIAN, Jonathan RAWSKI, and Brandon PRICKETT (2020), Probing RNN Encoder-decoder Generalization of Subregular Functions using Reduplication, *Proceedings of the Society for Computation in Linguistics*, 3(1):31–42.

Elissa L. NEWPORT and Richard N. ASLIN (2004), Learning at a Distance I. Statistical Learning of Non-adjacent Dependencies, *Cognitive psychology*, 48(2):127–162.

Charlie O'Hara and Caitlin Smith (2019), Computational Complexity and Sour-Grapes-like Patterns, in *Proceedings of the Annual Meetings on Phonology*, volume 7.

Brandon Prickett (2019), Learning Biases in Opaque Interactions, *Phonology*, 36(4):627–653.

Brandon Prickett, Aaron Traylor, and Joe Pater (2018), Seq2Seq Models with Dropout can Learn Generalizable Reduplication, in *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 93–100.

Alan Prince and Bruce Tesar (2004), Learning Phonotactic Distributions, *Constraints in phonological acquisition*, pp. 245–291.

Ezer Rasin and Roni Katzir (2016), On Evaluation Metrics in Optimality Theory, *Linguistic Inquiry*, 47(2):235–282.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen (2019), Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages, in *Proceedings of NAACL-HLT*, pp. 3532–3542.

Sharon Rose and Rachel Walker (2011), Harmony Systems, *The handbook of phonological theory*, 2:240–290.

DE Rumelhart and JL McClelland (1986), On Learning the Past Tenses of English Verbs, in JL McClelland and DE Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2: Psychological and Biological Models, pp. 216–271, The MIT Press.

Edward Sapir and Harry Hoijer (1967), *The Phonology and Morphology of the Navaho Language*, University of California Press.

Hava T. Siegelmann (1999), *Neural Networks and Analog Computation: Beyond the Turing Limit*, Springer Science & Business Media.

Caitlin Smith, Charlie O'Hara, Eric Rosen, and Paul Smolensky (2021), Emergent Gestural Scores in a Recurrent Neural Network Model of Vowel Harmony, *Proceedings of the Society for Computation in Linguistics*, 4(1):61–70.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014), Sequence to Sequence Learning with Neural Networks, in *Advances in Neural Information Processing Systems*, pp. 3104–3112.

David S. Touretzky (1989), Towards a Connectionist Phonology: The "Many Maps" Approach to Sequence Manipulation, in *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, pp. 188–195.

David S. Touretzky and Deirdre W. Wheeler (1990), A Computational Basis for Phonology, in *Advances in Neural Information Processing Systems*, pp. 372–379.

Paul TUPPER and Bobak SHAHRIARI (2016), Which Learning Algorithms Can Generalize Identity-Based Rules to Novel Inputs?, *arXiv preprint arXiv:1605.04002*.

Gail WEISS, Yoav GOLDBERG, and Eran YAHAV (2018), On the Practical Computational Power of Finite Precision RNNs for Language Recognition, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 740–745.

Gesche WESTPHAL-FITCH, Beatrice GIUSTOLISI, Carlo CECCHETTO, Jordan Scott MARTIN, and W. Tecumseh FITCH (2018), Artificial Grammar Learning Capabilities in a Visual Task Match Requirements for Linguistic Syntax, *Frontiers in psychology*, 9:1210.

Colin WILSON (2003), Analyzing Unbounded Spreading with Constraints: Marks, Targets, and Derivations, *Unpublished manuscript, University of California, Los Angeles*.

Brandon Prickett
0000-0001-9217-2130
bprickett@umass.edu

University of Massachusetts Amherst

Brandon Prickett (2021), *Modelling a subregular bias in phonological learning with Recurrent Neural Networks*, Journal of Language Modelling, 9(1):67–96
https://dx.doi.org/10.15398/jlm.v9i1.251