Capturing Phonotactic Learning Biases with a Simple RNN

Max Nelson
University of Massachusetts Amherst
manelson@umass.edu

Brandon Prickett
University of Massachusetts Amherst
bprickett@umass.edu

Joe Pater
University of Massachusetts Amherst
pater@linguist.umass.edu

Abstract

Recent artificial language learning experiments have used the six pattern types defined by Shepard et al. (1961) to explore what learning biases govern the acquisition of linguistic sound patterns (Moreton and Pertsova, 2016; Moreton et al., 2017). Moreton et al. (2017) went on to show that a Maximum Entropy Grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008) could correctly predict these biases, due to the structure of its phonological constraints. In this paper we show that a simple recurrent neural network can also predict the biases observed by Moreton et al. (2017). Not only does this demonstrate that the prespecified constraint set used by the Maximum Entropy model is not crucial for such a prediction, but because our learner is only exposed to positive evidence, unlike the model used by Moreton et al. (2017), its acquisition task is more analogous to the one performed by humans in the experiment.

1 Introduction

Shepard et al. (1961) defined six pattern types that represent every possible way to divide a three-feature stimulus space into two equal halves. They used these in tests of visual category learning to examine the relative ease of acquisition of the types. This inspired a large body of experimental and computational work, reviewed in Moreton et al. (2017).

Recent work on phonological learning (i.e., the acquisition of sound patterns in a language) has shown that the Shepard et al. pattern types are relevant in this domain as well, with interesting parallels and differences with the visual results (Moreton and Pertsova, 2014, 2016; Moreton et al., 2017). This paper will focus on one such study, in which Moreton et al. (2017) showed that in an artificial language learning experiment participants' pattern type preferences differed from Shepard et al.'s (1961) original ordering, but were predicted by a *Maximum Entropy* phonotactic learner (henceforth "MaxEnt" Goldwater and Johnson, 2003; Hayes and Wilson, 2008).

In this paper, we study a simple recurrent neural network (henceforth "RNN"; Jordan, 1986; Elman, 1990) that is designed to predict which sound comes next in a word, given the set of previous sounds in that word (for similar approaches, see Rodd, 1997; Mirea and Bicknell, 2019; Mayer and Nelson, 2020). We train this model on learning data that is structured identically to the data humans in the Moreton et al. (2017) experiment were exposed to. We then test it on the same kind of testing data that they used and find that the network captures the human learning just as well as the MaxEnt model, despite being given less linguistic structure *a priori*.

2 Background

The Shepard et al. (1961) pattern types can be exemplified with a stimulus space made up of every combination of the three binary features [±circle], [±black], and [±small]. There are eight possible stimuli in this space and six different ways of dividing the resulting stimulus space, exemplified in Figure (1). The patterns labeled "Type I" by Shepard et al. (1961) involve only a single feature—for example, "shapes that are [+black] are allowed in the pattern." A Type II pattern involves two features and a logical biconditional, such as "shapes that are either [+circle] and [+black] *or* [-circle] and [-black] are allowed." Types III-V use all three of the features, but a subset of the items can be grouped together using only a two-feature description. For example, "shapes that are [-black] and [-circle] *or* shapes that are [+black] and [+small] are allowed." Type VI also requires all three features, and needs all three to distinguish any allowable item from a banned one.

Moreton et al. (2017) implemented these six types as phonological patterns using features that described the sounds that a word contains. Specifi-

cally, they used words consisting of two consonant-vowel sequences in which the consonants could be [±voice] and [±coronal] and the vowels could be [±high] and [±back].[1] An example of the kind of the pattern used by Moreton et al. (2017) would be "only words with a [+voice] sound in the first consonant position are allowed." This would be a Type I pattern, since it only depends on the feature [voice] in the first consonant. In the experiment's testing phase, average participant accuracy on each pattern went in the order (from best to worst): I > IV > III > V > II > VI.

Moreton et al. (2017) simulate their experiments using a MaxEnt grammar that is provided with a constraint for every possible conjunction of the relevant features. For example, there would be a constraint $*[+voice]_{C1}$ that would be activated any time the first consonant in a word was [+voice], but there would also be a constraint like $*[+voice]_{C1}[-voice]_{C2}$ that only activated when the first consonant was [+voice] *and* the second consonant was [-voice]. They show that this model predicts the ordering I > III,IV > II,V > VI —with an early preference for pattern IV over pattern III that reverses late in learning, and an early preference for pattern V over pattern II that also eventually reverses.

A requirement of the MaxEnt model used by Moreton et al. (2017), as well as some past approaches to phonotactic learning with RNNs (e.g. Doucette, 2017), is that they are given the full probability distribution over all possible experiment stimuli in the training data (similar MaxEnt phonotactic models share this requirement; e.g., Hayes and Wilson, 2008). This means that they are exposed to stimuli that have a non-zero probability (i.e., positive evidence for the pattern) as well as stimuli have a probability of zero (i.e., negative evidence for the pattern). Since the humans participating in the Moreton et al. (2017) experiment were only given positive evidence (mimicking infants' exposure to real-world sound patterns), an ideal simulation of the experiment would limit its model to these items.

The RNN model we use here, described further in 3, is given less information than the model tested by Moreton et al. (2017). Its input features are only the four features necessary to describe each of the possible sounds and the model is only exposed to positive evidence for each pattern.

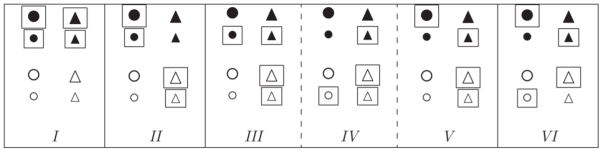

Figure 1: Illustration from Moreton et al. (2017) showing examples of each of Shepard et al.'s 1961 pattern types, using the binary features [±square], [±black], and [±small]. Shapes inside of squares are allowed in each type's example pattern.

3 RNNs as a Model of Phonotactic Learning

The models applied here are simple recurrent neural networks trained on a modified language modeling objective (Elman, 1990). RNN language models have been shown to be a viable method for modeling human phonotactic knowledge (Rodd, 1997; Doucette, 2017; Mirea and Bicknell, 2019; Mayer and Nelson, 2020), as well as phonological mappings (Hare, 1990; Gasser and Lee, 1992; Prickett, 2019). Unlike previous language modeling approaches, here we represent output phonemes as a vector of phonologically informed feature values. Both input and output segments are encoded as four-dimensional vectors with each position corresponding to one of the four phonological features used to characterize the patterns in Moreton et al. (2017) such that a 0 represents a negative feature value and a 1 represents a positive feature value. At each timestep the network predicts a four-dimensional output to which a sigmoid activation function is applied to yield the expected feature vector for the upcoming phoneme.

4 Methods

Following Moreton et al. (2017), experiments are run on an artificial language consisting of only 8 phonemes, 4 consonants [t, k, d, g] and the four vowels [i, u, æ, a]. Phonemes are represented with four features, *voice*, *coronal*, *high*, and *back* as shown in Table (1).

For each of the six Shepard type patterns, twenty-four different instantiations of that pattern are randomly generated using the feature set in Table (1).[2]

[1] These features were useful, since they represent standard features used in the phonological literature and correlate with articulatory characteristics in each of the sounds. Specifically, [voice] describes whether the vocal folds are vibrated when a consonant is pronounced, [coronal] refers to which part of the tongue a consonant is made with, and [high]/[back] describe the position of the tongue during a vowel.

[2] In this space, there are 8 features: two features for each

	voice	*coronal*	*high*	*back*
t	−	+	−	−
k	−	−	−	−
d	+	+	−	−
g	+	−	−	−
i	−	−	+	−
u	−	−	+	+
æ	−	−	−	−
a	−	−	−	+

Table 1: Feature specifications used for the 8 phoneme inventory

For each of those patterns a model is trained on just 32 randomly chosen pattern-conforming words. The test data is a forced-choice task in which a further 32 pattern-conforming words are paired with 32 pattern non-conforming words. This is the exact same procedure that Moreton et al. (2017) used when constructing the stimuli for their experiment.

For each train-test split models are trained for 200 epochs. Binary cross-entropy is optimized with adam (Kingma and Ba, 2014). Results below are reported for a model with a 32 dimensional hidden state.

Every 5 epochs the model is used to simulate the forced-choice test from the experiment. The forced-choice task is simulated using an adaptation of the Luce choice rule. For every pair of words w_1 and w_2, the network's loss on both words in calculated and then the probability of choosing either word is set to 1 minus its proportion of the total loss across both words.

$$p(w_1 \mid w_1, w_2) = 1 - \frac{\mathcal{L}(w_1)}{\mathcal{L}(w_1) + \mathcal{L}(w_2)} \quad (1)$$

5 Results

Figure (2) shows the average probability placed on the correct choice by epoch, averaged across all choices in ten randomizations of each pattern. Error bars represent 95% confidence intervals.

The predicted relative difficulty of the different Shepard type patterns aligns almost exactly with the predictions of Moreton et al. (2017)'s MaxEnt learner. The general ranking of the patterns from easiest to most difficult is I > III,IV > II,V > VI. Also in line with Moreton et al. (2017), there is an early bias for pattern V over pattern II and for pattern IV over III, both of which are lost later in learning. These results are also in line with Moreton et al.'s 2017 reported human accuracy ranking of I > IV > III > V > II > VI.

6 Discussion

We have shown that a simple RNN can model the biases observed in human phonotactic learning by Moreton et al. (2017). The MaxEnt model they used to model their results required a prespecified constraint set and negative evidence of the patterns to do so—both of which our RNN lacked.

Several avenues exist for future work. For example, one could test whether the structure our model *was* given at the start of learning (i.e., the phonologically informed features in its input and output) was necessary to capture the Shepard Type biases. Since the participants in the Moreton et al. (2017) experiment were all fluent speakers of English, phonemes could instead be represented using features based on the context the relevant phonemes appear in in English (for similar approaches to a different task, see Rodd, 1997; Mirea and Bicknell, 2019; Mayer and Nelson, 2020). Other biases observed in artificial language learning could also be tested, such as the *Intradimensional Bias* observed in phonotactic learning by Moreton (2012) or the *Identity Bias* observed by Gallagher (2013). Future could also explore how our model performs on real language data and see how its generalizations from such data compare to those of humans in experiments like Hayes and White (2013) and Berent et al. (2002). Additionally, our model and results could be compared to the neural modeling of visual category learning (e.g. Kurtz, 2007) to see if it can be extended to yield the I > II > III, IV, V > VI order of the original Shepard et al. experiments.

Past computational approaches to deriving the observed Shepard type biases, including those based on MaxEnt models and other neural methods often provided the learner with explicit negative evidence (e.g. Kruschke, 1992; Moreton et al., 2017, although, see Kurtz, 2007). The language modeling approach taken here makes use of indirect negative evidence that is likely much more closely aligned with the kind of evidence used by real language learners (Clark and Lappin, 2009). The lan-

segment, and 4 segments. Each pattern uses 3 of them to define a Shepard type, so for any given pattern, there will be several irrelevant features. For example, if a pattern bans all words without a [+voice] value for their first consonant, the features [coronal], [high], and [back] (as well as [voice] in the second consonant) will be unnecessary for determining whether a word is allowed or not.

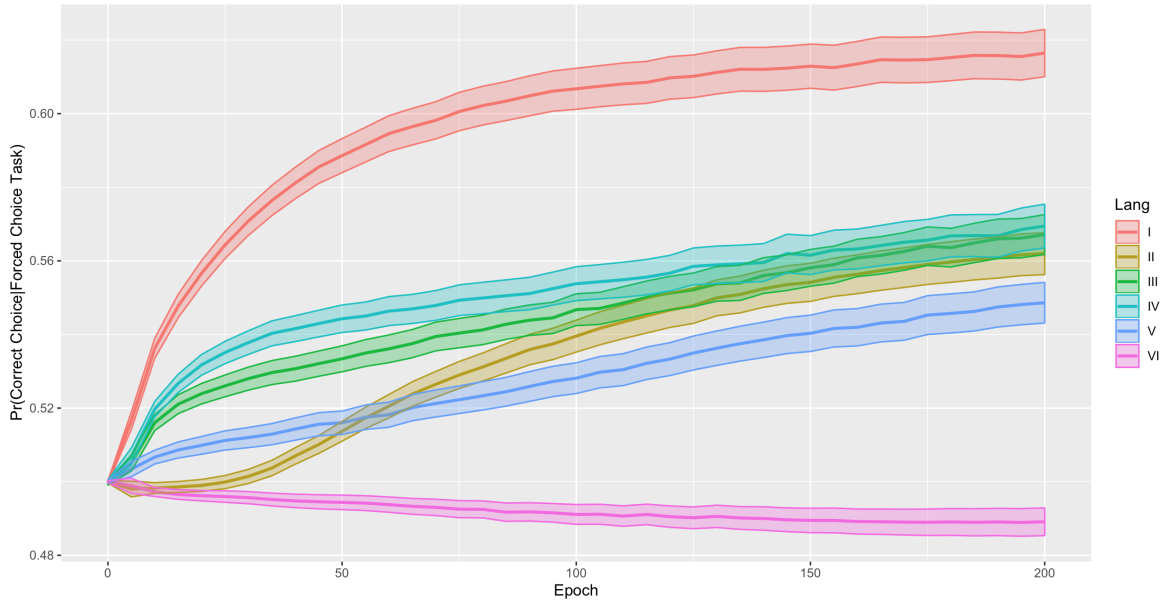

Figure 2: Average probability of the correct (in-language) choice in the forced-choice task by epoch for six Shepard type languages

guage modeling objective also bears similarities to the types of predictive processing that has been observed both in the phonological and syntactic domains (see, e.g., Grosjean, 1980). The relationship between language modeling and phonotactic learning, both in terms of supervision and predictive processing, is a promising area for future work.

Finally, Moreton et al. (2017) show that the behavior of their MaxEnt model is a natural consequence of their proposed conjunctive constraint schema. The neural models tested here are not explicitly provided with a similarly structured learning space and exhibit nearly identical biases. Future work should examine the weight space of the network to see if analogs to Moreton et al. (2017)'s constraints can be identified, or should otherwise aim to explain why these biasing are emerging in our recurrent networks.

These results are the latest in a line of work (Rodd, 1997; Doucette, 2017; Mirea and Bicknell, 2019; Mayer and Nelson, 2020) suggesting that RNNs are a viable model for human phonotactic learning. They also show that the kind of *a priori* constraint set given to the MaxEnt model used by Moreton et al. (2017) is not necessary for capturing the biases observed in their experiment—meaning that such constraints may not be an innate feature of human learning.

References

Iris Berent, Gary Marcus, Joseph Shimron, and Adamantios I. Gafos. 2002. The scope of linguistic generalizations: Evidence from Hebrew word formation. *Cognition*, 83(2):113–139.

Alexander Clark and Shalom Lappin. 2009. Another look at indirect negative evidence. In *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 26–33.

Amanda Doucette. 2017. Inherent Biases of Recurrent Neural Networks for Phonological Assimilation and Dissimilation. *arXiv preprint arXiv:1702.07324*.

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

Gillian Gallagher. 2013. Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, 30(2):253–295.

Michael Gasser and Chan-Do Lee. 1992. Networks that learn about phonological feature persistence. In *Connectionist Natural Language Processing*, pages 349–362. Springer.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111120.

François Grosjean. 1980. Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, 28(4):267–283.

Mary Hare. 1990. The role of trigger-target similarity in the vowel harmony process. In *Annual Meeting of the Berkeley Linguistics Society*, volume 16, pages 140–152.

Bruce Hayes and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, 44(1):45–75.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Michael I Jordan. 1986. Serial order: A parallel distributed processing approach (Tech. Rep. No. 8604). Technical report, University of California, Institute for Cognitive Science, San Diego, CA.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John K Kruschke. 1992. Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22.

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

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics*, 3(1):149–159.

Nicole Mirea and Klinton Bicknell. 2019. Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1595–1605.

Elliott Moreton. 2012. Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, 67(1):165–183.

Elliott Moreton, Joe Pater, and Katya Pertsova. 2017. Phonological Concept Learning. *Cognitive science*, 41(1):4–69.

Elliott Moreton and Katya Pertsova. 2014. Pastry phonotactics: Is phonological learning special. In *Proceedings of the 43rd Annual Meeting of the Northeast Linguistic Society, City University of New York*, pages 1–14. Graduate Linguistics Students' Association Amherst, MA.

Elliott Moreton and Katya Pertsova. 2016. Implicit and explicit processes in phonotactic learning. In *Proceedings of the 40th Boston University Conference on Language Development, Somerville, Mass., pp. TBA. Cascadilla*.

Brandon Prickett. 2019. Learning biases in opaque interactions. *Phonology*, 36(4):627–653.

Jennifer Rodd. 1997. Recurrent neural-network learning of phonological regularities in Turkish. *CoNLL97: Computational Natural Language Learning*.

Roger N. Shepard, Carl I. Hovland, and Herbert M. Jenkins. 1961. Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1.