Explaining differences between phonotactic learning biases in the lab and typological trends using Probabilistic Feature Attention

Brandon Prickett
University of Massachusetts Amherst
bprickett@umass.edu

1 Introduction

A primary goal of linguistic theory is to explain why certain kinds of languages are underattested. One methodology that has had success in explaining phonological typology has been artificial language learning, in which participants are trained for a short period of time on a synthetic language that was designed to test the learnability of a particular kind of pattern (for a review of this literature, see Moreton and Pater, 2012a,b). Often, the goal of these experiments is to see if participants' learning biases in the lab might explain typology by showing that underattested languages are more difficult to acquire (see, e.g., Wilson, 2006; Finley, 2008; Glewwe, 2019).

However, learning biases seen in an experimental setting do not always match typological trends. Moreton and Pertsova (2014) implemented a set of patterns introduced by Shepard et al. (1961, henceforth, *Shepard Types*) as phonotactic restrictions and taught them to participants in an artificial language learning experiment. They found that participants' preferred patterns failed to mirror typological trends in a database of attested phonological generalizations (Mielke, 2008).

Here, I model the acquisition of phonotactic patterns that align with the six Shepard Types tested by Moreton and Pertsova (2014) using a maximum entropy phonotactic grammar (Hayes and Wilson, 2008; Moreton et al., 2017) equipped with Probabilistic Feature Attention (Prickett, 2023). This model predicts the biases seen in Moreton and Pertsova (2014)'s experimental results early in learning, but by the end of learning reflects the trends present in phonological typology. These results could help explain the differences observed by Moreton and Pertsova (2014) between artificial language learning and typology, since the latter could be shaped by more long-term learning biases.

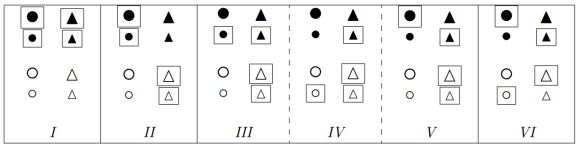

Figure 1: Shepard Type examples using the features [±black], [±circle], and [±large]. Boxes around shapes show how stimuli could be divided up in each type. Taken from Moreton et al. (2017).

2 Background

2.1 Shepard Types

Shepard et al. (1961) found that humans were biased toward certain kinds of patterns when learning in an experimental setting. They used patterns involving 8 stimuli, where each stimulus could be uniquely identified with three features. The shapes in Figure 1 show an example of such a stimulus space. They found that out of the six possible ways of dividing up the space into two equally sized groups, their participants learned some divisions more quickly than others. The roman numerals in Figure 1 show the relative ease with which each type was learned in their original experiments (with lower numbers applied to easier Types and dotted lines between Types representing inconsistent/marginal differences in learnability). For a review of the literature on Shepard Types for non-linguistic patterns, see Kurtz et al. (2013).

Moreton and Pertsova (2014) implemented the Shepard Types as phonotactic patterns (where the three features were phonological and the stimuli were words). Their results showed that in this context, the Shepard Types were learned in the order (from easiest to most difficult): I, IV, III, V, II, and VI. However, when Moreton and Pertsova (2014) analyzed a database of phonological patterns, assigning as many patterns as they could to each of the Shepard Types, they found that the typological frequency of the Types roughly mirrored the origi-

Segment	[labial]	[continuant]	[voice]
p	+	-	-
b	+	-	+
f	+	+	-
v	+	+	+
t	-	-	-
d	-	-	+
s	-	+	-
z	-	+	+

Table 1: Features and segments used for all simulations presented here.

nal ordering found by Shepard et al. (1961): I, II, III, IV/V, and VI.

2.2 Probabilistic Feature Attention

Prickett (2023) proposed Probabilistic Feature Attention (henceforth, *PFA*) as a way to model certain kinds of uncertainty that likely exist in the process of phonological acquisition. PFA introduces noise into a learning model's training data by making certain segments temporarily ambiguous with one another and is based on a regularization technique from the machine learning literature called *dropout* (Srivastava et al., 2014). This ambiguity is based on the features used to represent the segments, with the model distributing its attention (Nosofsky, 1986) to these features probabilistically and resampling which features are attended to on each learning update.[1]

For example, imagine a phonotactic pattern using the segments and features in Table 1. If the model attended to the feature [continuant], but not [voice], the difference between [t] and [s] would be preserved, but the model would treat [t] and [d] identically. If the model was learning a pattern in which voiceless sounds were grammatical and voiced sounds were not, any learning update in which [voice] was not attended to would fail to push the learner in the correct direction.

Prickett (2023) paired PFA with a maximum entropy phonotactic learner (Hayes and Wilson, 2008) with a conjunctive constraint schema (Moreton et al., 2017) and successfully modeled a number of artificial language learning experiments. Those results demonstrated that some relevant features being attended to while others are not can push the model to generalize and learn in unexpected ways. This altered learning and generalization mirrored the human behavior in the relevant experiments.

[1]Note that this ambiguity could arise from a number of factors in real phonological acquisition, such as misperception (Bailey and Hahn, 2005) or constraints on memory (Gathercole and Adams, 1993).

3 Methods

The results presented here were found using the software published in the supplementary materials included with Prickett (2023), which implements a maximum entropy phonotactic grammar and trains it with batch gradient descent paired with PFA. The hyperparameter values that were used for these results were a learning rate of .05 and an attention probability of .25. These were chosen after a short amount of piloting, with a full grid search of these values being left to future work.

Constraints representing every possible combination of the features in Table 1 were used (following Moreton et al., 2017). This included constraints with a single valued feature (e.g., *[+voice]), constraints with two valued features (e.g., *[+voice, +continuant]), and constraints with three valued features (e.g., *[+voice, +continuant, -labial]). Constraints with a single feature were always violated by half of the possible segments (e.g., [b, v, d, z]), constraints with two features were always violated by two segments (e.g., [v, z]), and constraints with three features were always violated by a single segment (e.g., [z]).

Six 'languages' (one for each Shepard Type) were implemented using 'words' that were only a single segment long. In the training data for each language, four of the words had a probability of 1 and four had a probability of 0 (representing grammatical and ungrammatical words, respectively). The model was tested in 30 separate runs for each language, since PFA introduces variability into the learning process. This ensured that results were representative of the model's average behavior, and not the random choice of feature attention in a single run.

4 Results

Figure 2 shows the average accuracy for the model with PFA on each pattern. The model's initial ordering of Shepard Types matches the performance observed by Moreton and Pertsova (2014) in their experiment: I, IV, III, V, II, and VI. However, later in learning, the ordering of the patterns mirrors the typological trends observed by Moreton and Pertsova (2014), instead, with Type II crucially having a higher accuracy than III, IV, or V. Note

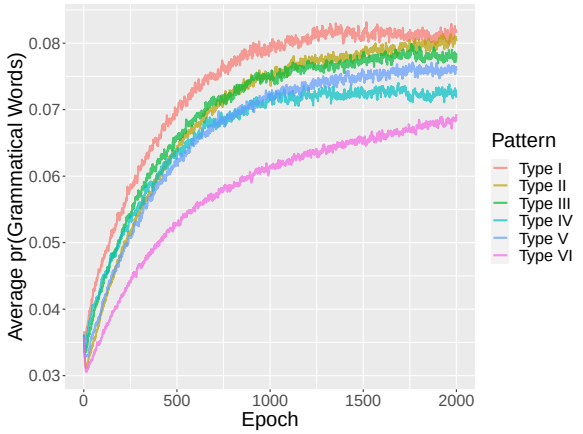

Figure 2: Probability of grammatical words in each pattern, according to the model at each epoch. Results are averaged over 30 separate runs per pattern.

that the ordering of Types IV and V does change toward the end of acquisition, but the relative ordering of these Types in the typological study was also inconsistent.

5 Discussion

5.1 Why does the model capture these biases?

The relative ordering of types that is present early on in the model's learning matches the expected behavir of this kind of maximum entropy learner. The reason that these biases exist in the model is because of the structure of its constraint set and the nature of gradient-based learning algorithms. For an in-depth explanation for how a conjunctive constraint schema combined with gradient descent predicts this ordering of Shepard Types, see Moreton et al. (2017).

But why does PFA cause the model to change its relative ordering of Shepard Types later in learning? The more features that are relevant to a pattern, the more opportunities PFA has to obscure that pattern over the course of learning (for more on this effect, see Prickett, 2023, §4.3). Type II patterns only involve two features, while Types IV and V both involve three. For Types IV and V, all three features must be attended to for a learning update to push the model in the correct direction. But in Type II, only the two relevant features have to be attended to for the model to move its weights in the correct direction. This effect of PFA compounds as learning continues, making IV and V ultimately more difficult to learn.

5.2 Future Work

The relationship between phonological learning in the lab and phonological typology in the real world is still largely an open question. Many factors could drive differences between biases seen in artificial language learning and real-world typology, such as the effect of sleep on acquisition (see e.g., St Clair and Monaghan, 2008), the pressures caused by the iterative and interactive nature of language learning (see e.g., Hughto, 2020), and phonetically driven channel bias (see e.g., Ohala, 2014). The results presented here offer an explanation for one particular mismatch between observed learning biases and the frequency of attested patterns, but future work should explore how PFA might interact with these other phenomena.

Future work should also explore whether other models of phonological learning can explain the results in Moreton and Pertsova (2014). A maximum entropy model that uses a conjunctive constraint schema will always predict the ordering of Shepard Types seen in Moreton and Pertsova (2014)'s experiment unless additional mechanisms are added to it. But other approaches to phonotactic learning, such as induced constraints (see, e.g. Hayes and Wilson, 2008), expectation-driven learning algorithms (Jarosz, 2015), or neural networks (see, e.g. Mayer and Nelson, 2020) could all be tested on these same patterns.

More typological work could also illuminate future directions for this kind of research. Moreton and Pertsova (2014) used patterns across two segments in their experiment, but only had access to single-segment patterns in the database they used to calculate typological frequencies (Mielke, 2008). The simulations presented here used single-segment patterns as well, but PFA can be used with multi-segment sequences (Prickett, 2023) and if future work found a different typological distribution for patterns involving two segments, testing the model on that kind of pattern could be useful.

5.3 Conclusions

While the goal of artificial language learning is usually to explain some kind of typological trend, Moreton and Pertsova (2014) found distinct differences between learning observed in the lab and the frequency of certain patterns in phonological typology. A model with PFA, an independently motivated mechanism (Prickett, 2023), matches Moreton and Pertsova (2014)'s experimental results early

in learning, but mirrors typological trends later in acquisition, providing a potential explanation for the mismatch observed by Moreton and Pertsova (2014).

References

Todd M. Bailey and Ulrike Hahn. 2005. Phoneme similarity and confusability. *Journal of Memory and Language*, 52(3):339–362.

Sara Finley. 2008. *Formal and Cognitive Restrictions on Vowel Harmony*. PhD Thesis, Johns Hopkins University.

Susan E. Gathercole and Anne-Marie Adams. 1993. Phonological working memory in very young children. *Developmental Psychology*, 29(4):770–778. Place: US Publisher: American Psychological Association.

Eleanor Glewwe. 2019. *Bias in Phonotactic Learning: Experimental Studies of Phonotactic Implicationals*. PhD Thesis, UCLA.

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

Coral Hughto. 2020. *Emergent typological effects of agent-based learning models in maximum entropy grammar*. Ph.D. thesis, University of Massachusetts Amherst.

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

Kenneth J Kurtz, Kimery R Levering, Roger D Stanton, Joshua Romero, and Steven N Morris. 2013. Human learning of elemental category structures: revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):552.

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

Jeff Mielke. 2008. *The emergence of distinctive features*. Oxford University Press.

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

Elliott Moreton and Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part II: Substance. *Language and linguistics compass*, 6(11):702–718.

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.

Robert M. Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39.

John Ohala. 2014. The phonetics of sound change. In *Historical linguistics*, pages 237–278. Routledge.

Brandon Prickett. 2023. Probabilistic feature attention as an alternative to variables in phonotactic learning. *Linguistic Inquiry*, 54(2):219–249.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Michelle C St Clair and Padraic Monaghan. 2008. Language abstraction: Consolidation of language structure during sleep. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982.