Simulating Variability-Induced Learning Biases Using Maxent Grammars

Overview

Recent work by Do and colleagues (e.g., Do and Mooney, 2022; Mooney and Do, 2018) suggests that learning biases towards substantively grounded, phonetically natural patterns may be more likely to be revealed when the learner is faced with variability. For example, child learners showed a bias towards vowel harmony (a phonetically grounded phonological pattern) over vowel disharmony (a less phonetically natural pattern) when the children were given a variable pattern (e.g., most, but not all items displayed the target phonological pattern) (Do and Mooney, 2022). Some learning biases may be more sensitive to variability if learners rely more on prior or substantively grounded biases in the face of ambiguous or conflicting evidence of the type that occurs in variable phonological patterns. However, it is unclear whether or how such a reliance on substantive grounding might be implemented formally, such as with Maxent learning models of Harmonic Grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008) for a phonetically motivated pattern like vowel harmony. The present study makes use of the Maxent Grammar Learning Tool (Hayes et al., 2009) to explore whether and when learning biases are more likely to emerge in a variable vs. categorical system. The results of three sets of simulations demonstrate that Maxent Grammars can simulate a stronger learning bias under variable conditions, but only if the bias towards harmony is put into the initial prior weight.

Simulations

The simulations were based on a simple grammar with two markedness constraints: AGREE (e.g., *[αF][βF]) which induces vowel harmony, and DISAGREE (e.g., *[αF][αF]) which induces disharmony. The grammar was evaluated based on two sets of bisyllabic candidates. The first evaluation compared a word with [+F][+F] vowels with a word with [–F][+F] vowels, where [+F][+F] satisfies AGREE, but violates DISAGREE. The second evaluation compared a word with [–F][–F] vowels with a word with [+F][–F] vowels, where [–F][–F] satisfies AGREE, but violates DISAGREE. The input to the categorical grammars consisted of 12 items for each of the dominant inputs: (e.g., 12 [+F][+F] and 12 [–F][–F] items in the Categorical Harmony grammar). The input to the variable grammars consisted of 9 items for each of the dominant inputs: (e.g., 9 [+F][+F], 3 [–F][+F], 9 [–F][–F], and 3 [+F][–F] items in the Variable Harmony grammar). A sample table for Variable Disharmony simulations is shown in Table 1.

Table 1: Sample Input Table for Categorical Disharmony Simulations.

Input	**Candidate**	**n**	**AGREE**	**DISAGREE**
++ vs. –+	[+F][+F]	3	0	1
	[–F][+F]	9	1	0
– – vs. +–	[-F][-F]	3	0	1
	[+F][-F]	9	1	0

Learning was simulated using the MaxEnt Grammar Learning Tool (Hayes et al., 2009). This tool has two user-defined parameters: σ^2 (a parameter that roughly corresponds to a learning weight, where lower values keep the final weight closer to the initial weight), and μ (a parameter that roughly corresponds to the initial weights). Varying these user-defined parameters for different constraints allows the researcher to simulate learning biases. For example, White (2017) induced bias by manipulating the initial value of μ, based on the assumption that learners came into the experiment with different levels of prior knowledge of the constraints. Wilson (2006) and Finley (2022) induced a bias by manipulating the value of σ^2, with the assumption that learners had no prior knowledge of the constraints going into the study, but had a bias for learning different constraint weights. A third possibility is that learners have a bias on both σ^2 and μ. In this situation, the learners have a bias for a specific constraint, and a bias against (or for) changing the weight of that constraint. The present study created simulations based on these three types of biases, and two control (non-biased) simulations. The first simulation included a bias in μ for Agree, which was set to 1.41, based on the learned weight of an unbiased simulation (following White 2017). The value for μ was set to 0 for Disagree. The value of σ^2 was 0.6 for all conditions, following White (2017). The second set of simulations set σ^2 at 0 for both constraints but lowered the value of μ to 0.1 for Disagree, creating a bias against changing the 0 weight. The third set of simulations used a biased value of Agree for both σ^2 (1.41) and μ (0.1), with the logic that the lower value of μ would bias the learner against changing the initial weight of Agree. The two unbiased control simulations set σ^2 at 0 and μ at either 0.6 or 0.1 for both constraints.

The results of the simulations are shown in Table 2, with the estimated proportion of the dominant value displayed. In all biased simulations, performance was better for the harmony languages. However, the bias was greater for the categorical conditions when the bias was placed on σ^2 alone: a 19-point difference (0.80 vs. 0.61) in the categorical simulations, but only a 10-point difference (0.66 vs. 0.56) in the variable simulations. When the bias was placed on μ, the difference between harmony and disharmony was bigger in the variable conditions. This difference was most apparent in the set of simulations where the bias was on both μ and σ^2. This suggests that biases are more likely to emerge under variability when the bias is on the initial weight.

Table 2: Summary and Results of Simulations. Proportions indicate proportion dominant output.

Bias	**Agree Parameters**		**Disagree Parameters**		**Results: Harmony Categorical**	**Results: Disharmony Categorical**	**Results: Harmony Variable**	**Results: Disharmony Variable**
	μ	σ^2	μ	σ^2				
μ	1.41	0.6	0	0.6	0.90	0.80	0.77	0.62
σ^2	0	0.6	0	0.1	0.80	0.61	0.66	0.56
μ, σ^2	1.41	0.1	0	0.6	0.83	0.72	0.77	0.56
none	0	0.6	0	0.6	0.80	0.80	0.66	0.66
none	0	0.1	0	0.1	0.61	0.61	0.56	0.56

Significance

Simulating the results of human learning with the Maxent Grammar Tool may provide insight into the mixed results of artificial language learning studies testing for substantive biases (Moreton and Pater, 2012). All simulations showed better performance for harmony over disharmony, even with categorical training data. This predicts that experimental learning settings might occasionally show differences between categorical grounded and ungrounded phonological patterns when there is a strong bias and a sensitive test of learnability. In variable learning conditions, the training data is harder to master, as it is more ambiguous. In this case, learners must hold on to their initial biases about the constraints (if they have them) to make sense of the pattern. Thus, when the bias is in the initial weight, a learning bias may be more likely to be revealed under variable conditions.