Simulating Variability-Induced Learning Biases Using MaxEnt Grammars

Sara Finley

Pacific Lutheran University

1 Background

Artificial language learning studies have proven to be a useful tool to understand the nature of phonological processes, particularly how learning biases may shape common (and uncommon) phonological patterns. However, much of the research that has explored whether learners are biased to phonetically grounded patterns over ungrounded patterns has yielded inconsistent, or mixed results (Moreton and Pater 2012). This is particularly true for comparisons between vowel harmony and vowel disharmony. In vowel harmony, vowels within a particular phonological domain must share the same phonological feature. In vowel disharmony, adjacent vowels must disagree for a particular phonological feature. Vowel harmony is typologically robust across language families and shows strong phonetic grounding (Ohala 1994). Both articulatory (e.g., coarticulation) and perceptual pressures (e.g., increasing redundancies and phonetic licensing) have been cited as phonetic motivators for vowel harmony, and may potentially explain why vowel harmony is such a common process cross-linguistically. Vowel disharmony, on the other hand, has only one advantage of creating distinction between adjacent vowels, but goes against coarticulatory pressures. Because local harmony rules that apply between two vowels are logically similar between harmony and disharmony processes, if learners show a bias for harmony over disharmony, it suggests that learners are sensitive to phonetic grounding in learning (Martin and Peperkamp 2020). However, several artificial language learning studies that directly compared vowel harmony to vowel disharmony showed no significant differences between vowel harmony and vowel disharmony (Pycha et al. 2003; Skoruppa and Peperkamp 2011). Martin and White (2021) showed a significant advantage to vowel harmony, but only in conditions that extended beyond two syllables, where the constraints that govern harmony and disharmony vary; since harmony can apply iteratively but disharmony cannot, it makes sense that learners should show a preference to vowel harmony under iterative conditions. Other studies have shown a bias for vowel harmony over vowel disharmony, but only under variable conditions (Huang and Do 2023). When the vowel (dis)harmony rule was presented as a categorical rule that applied 100% of the time, participants showed no difference in learning, but when the rule was presented as a variable rule (e.g., harmony in 75% of the trials), participants were more likely to accept a majority harmonic pattern than a majority disharmonic one. In a study with child language learners, participants seemed to reverse the majority disharmony pattern to vowel harmony when exposed to a variable pattern (Do and Mooney 2022). This suggests that exposure to some vowel harmony can help trigger the bias towards vowel harmony over disharmony, but there is no specific bias towards categorical vowel harmony.

A question that remains is why variability helps induce a bias towards vowel harmony over disharmony. Previous research has shown a general processing and perceptual bias towards harmony (Kimper 2017), and vowel harmony is phonetically grounded (Ohala 1994). One possibility is the bias towards vowel harmony is relatively small, since vowel disharmony must be a learnable pattern, as it is typologically viable. This small bias means that under favorable but noisy learning conditions (as in an artificial language learning experiment with categorical data), there is no clear preference for vowel harmony over vowel disharmony. However, when variability is introduced, the learning problem becomes more challenging. The small bias for harmony may help the harmonic words to stand out, which will favor conditions with a harmony majority,

* The author would like to thank Youngah Do, anonymous reviewers, and the organizers and audience of AMP 2023 for their helpful feedback and support. I assume responsibility for any and all errors.

Proceedings of *AMP 2023/2024*

but disfavor conditions with a disharmony majority. The goal of the present study is to test this hypothesis through a simulation of both categorical and variable learning conditions using MaxEnt Harmonic Grammar (Goldwater and Johnson 2003; Hayes and Wilson 2008; Hayes, Wilson, and George 2009). In a MaxEnt grammar, it is possible to manipulate the learning rate, biases towards specific constraints, as well as the input and candidates. These manipulations make it possible to understand how learning biases and training input interact in learning. The simulations in this study show that greater learnability of vowel harmony can be observed under variable conditions when there is a prior bias towards the harmony-inducing constraint.

2 Method

The MaxEnt Grammar Learning Tool (Hayes, Wilson, and George 2009) simulates learning condition using the maximum entropy learning algorithm (Goldwater and Johnson 2003). For more details on the mathematical underpinnings of this model, the reader is invited to read Goldwater and Johnson (2003) as well as Hayes and Wilson (2008) for an application of the model to vowel harmony. For details on the mathematics of the model as applied to artificial language learning studies, the reader is also invited to read Wilson (2006) and White (2017). The MaxEnt learning algorithm essentially finds the optimal weights for a set of constraints, given a specific input set that contains the input-candidate mappings and their constraint violations. In order to avoid overfitting, the model makes use of a Gaussian prior for each constraint. This prior can serve as a bias for both the initial weights, as well as create penalties for changing the weights too quickly. The MaxEnt Grammar Learning Tool allows the user to provide the program with a list of inputs and their subsequent constraint violations. The user also must specify two learning parameters: μ, a parameter roughly corresponding to starting weights of the constraints, and σ^2, a parameter that sets the shape of the prior distribution, such that lower values require more data to change the constraint weight. The output of the program includes the learned weights for each constraint, as well as the associated probabilities for each candidate. All files associated with the simulations can be found at: https://osf.io/pyrxq/.

2.1 *Harmony Constraints and Input* In this model, I assume a very simple grammar, where all words are bisyllabic and all vowels are either [+F] or [–F]. The model also assumes that harmony can be induced by the AGREE constraint, which is violated by either [+F][–F], or [–F][+F] (i.e., when the vowels disagree in the harmonic feature). Disharmony is induced by the DISAGREE constraint, which is violated by either [+F][+F], or [–F][–F] (i.e., when the vowels agree in the harmonic feature). This creates a complementary distribution; when AGREE is satisfied, DISAGREE is violated, and vice versa. While there are other possible constraints to induce harmony and disharmony that may capture more complex vowel harmony patterns better (Walker 2012; Finley 2024), the AGREE/DISAGREE constraints are regularly used to account for naturally occurring vowel harmony patterns (e.g., (Baković 2000). These constraints are also sufficient to provide insights into the relevant descriptive driving force of vowel harmony and disharmony (agreement of vowels).

The input to the simulation contained 24 total items across four types of bisyllabic words: [+F][–F], [+F][+F], [–F][–F], and [–F][+F]. These were divided into two inputs, one pitting [–F][–F] against [–F][+F], and another pitting [+F][+F] against [+F][–F]. In the categorical conditions, the n for non-conforming items was 0 (e.g., in the Categorical Harmonic condition there were 12 [–F][–F] items, and 12 [+F][–F] items). The variable conditions had nine of the majority pattern, and three of the minority pattern (e.g., in the Variable Vowel Harmonic condition there were nine [–F][–F] items, and nine [+F][–F] items). This is spelled out in Table 1 below.

Candidates	**Categorical Harmonic**	**Categorial Disharmonic**	**Variable Harmonic**	**Variable Disharmonic**
[–F][–F]	12	0	9	3
[–F][+F]	0	12	3	9
[+F][+F]	12	0	9	3
[+F][–F]	0	12	3	9

Table 1: Simulation Input

2.2 *Learning Parameters* In addition to setting the input and the constraint set, the user must also set values for σ^2 and μ, which as noted above, can be used to induce a bias on the prior weight and reweighting of a specific constraint. The values based on these parameters have been used to induce biases in learning (Wilson 2006; White 2017; Finley 2022). Wilson (2006) placed the perceptually based bias towards velar palatalization on σ^2, and Finley (2022) used σ^2 to model perceptual similarity of consonants. White (2017) modeled the bias based on confusability data that changed the values of μ but kept the value of σ^2 constant. While Finley (2023) was able to model a bias for vowel harmony in exceptions without any specific bias for the harmony constraint, there was no proposed a prior phonetic or cognitive bias towards harmony, but a formal property of the constraint interaction.

Because it is possible to induce a learning bias on by changing either (or both) σ^2 and μ, I ran multiple simulations with the biases on different parameters in order to test whether the location of the bias might make a difference to learning, and the bias for vowel harmony over disharmony under different levels of variability in the input. For example, if the bias is placed on σ^2, it may suggest that learners have no pre-determined knowledge of AGREE over DISAGREE, but that they are able to reweight the AGREE constraint with less information than the DISAGREE constraint. If the bias is placed on μ, it could suggest that learners start with some knowledge of the AGREE constraint. However, it is also important to note that both μ and σ^2 are parameters used by the model to determine the probability of the data given the evidence, and both working together can influence the final weighting of constraints and probabilities assigned to each candidate. It is important to note, however, that while creating multiple simulations can help to distinguish between biasing methods, there are numerous ways to implement each strategy. This large degree of freedom means that the results should not be taken as a blanket result for all possible implementations.

I created several different simulations each with different values for σ^2 and μ. The first simulation placed the harmony bias on μ. I followed White's (2017) procedure of using the output of an unbiased run the MaxEnt Grammar tool on the biased items (e.g., harmonic items) to get the input weight. Doing this yielded a value for μ of -1.41 for AGREE and 0 for DISAGREE. Following White (2017), I kept the value of σ^2 at 0.6 for both constraints. The second simulation set the harmony bias on σ^2. I followed Wilson's (2006) procedure and set μ to 0 for both constraints, σ^2 to 0.1 for DISAGREE 0.6 for AGREE, which should have the effect of making it easier to reweight AGREE over DISAGREE. A third simulation was also created that carried the bias on both σ^2 and μ, and combined both strategies, with μ set to -1.41 for AGREE, and σ^2 set to 0.1 for AGREE, and 0.6 for DISAGREE. This was meant to have the effect of keeping the weight on AGREE, more evidence in favor of DISAGREE. The fourth simulation served as a control, and set μ to 0 for both constraints, and σ^2 to 0.6 for both constraints. We expect that there should be no bias for vowel harmony over disharmony in this simulation, but that the model will show a stronger preference for the majority pattern when it is categorical.

3 Results

3.1 *Bias on μ* When the bias was placed on μ (-1.41 for AGREE), the model showed a small difference in learning for categorical harmony (90%) over categorical disharmony (80%) but showed a bigger difference between learning variable harmony (77%) over variable disharmony (62%), as shown in Table 2. These findings generally simulate the results of the adult experimental literature. The difference between categorical harmony and disharmony is just 10%, while the difference between variable harmony and disharmony is 15%. With noisy human data, it is likely that a 10% difference between categorical harmony and categorical disharmony would only reach statistical significance under very ideal conditions or with a very large sample size.

Candidates	Categorical Harmonic	Categorial Disharmonic	Variable Harmonic	Variable Disharmonic
[-F][–F]	0.90	0.20	0.77	0.38
[-F][+F]	0.10	0.80	0.23	0.62
[+F][+F]	0.90	0.20	0.77	0.38
[+F][-F]	0.10	0.80	0.23	0.62

Table 2: Simulation Output, Bias on μ

3.2 *Bias on σ^2* When the bias was placed on σ^2 (0.1 for DISAGREE, and 0.6 for AGREE) the learner showed higher rates of learning categorical vowel harmony (80%) over categorical disharmony (61%), a difference of 19%, as shown in Table 3. The difference between harmony and disharmony was smaller for the variable conditions (66% for vowel harmony and 56% for vowel disharmony, for a difference of 10%). While the difference between harmony and disharmony is smaller for the variable than the categorical conditions, it is important to note that the low performance of the variable disharmony condition suggests that it would not be learnable at a rate significantly higher than chance. This is the opposite pattern of the human data, where the differences were bigger under variable conditions.

Candidates	Categorical Harmonic	Categorial Disharmonic	Variable Harmonic	Variable Disharmonic
[–F][–F]	0.80	0.39	0.66	0.44
[–F][+F]	0.20	0.61	0.34	0.56
[+F][+F]	0.80	0.39	0.66	0.44
[+F][–F]	0.20	0.61	0.34	0.56

Table 3: Simulation Output, Bias on σ^2

3.3 *Bias on μ and σ^2* When the bias was placed on both μ and σ^2, the model showed better learning for harmony over disharmony in both the categorical (83% vs. 72%), and the variable conditions (77% vs. 56%), as shown in Table 4. However, the difference was greater for the variable condition (11% vs. 21%). In addition, the the variable disharmony condition was barely above 50%, suggesting a general lack of preference for disharmony when it was the majority pattern. This simulation may be the closest one to the human data, where the biggest difference between harmony and disharmony is shown in the variable conditions, and the model failed to learn the variable disharmonic pattern.

Candidates	Categorical Harmonic	Categorial Disharmonic	Variable Harmonic	Variable Disharmonic
[-F][–F]	0.83	0.28	0.77	0.44
[-F][+F]	0.17	0.72	0.23	0.56
[+F][+F]	0.83	0.28	0.77	0.44
[+F][-F]	0.17	0.72	0.23	0.56

Table 4: Simulation Output, Bias on μ and σ^2

3.4 *No Bias* When there was no bias placed on any constraint, the model learned harmony and disharmony at the same rate, but at a lower rate than in the variable conditions. The model learned the categorical harmony and disharmony at a rate of 80% each for harmony and disharmony, and the variable rules at a rate of 66% each for both rules. This simulates a model where a learner roughly probability matches when faced with variable data.

Candidates	Categorical Harmonic	Categorial Disharmonic	Variable Harmonic	Variable Disharmonic
[-F][–F]	0.80	0.20	0.66	0.34
[-F][+F]	0.20	0.80	0.34	0.66
[+F][+F]	0.80	0.20	0.66	0.34
[+F][-F]	0.20	0.80	0.34	0.66

Table 4: Simulation Output, No Bias

4 Discussion and Conclusion

All of the biased parameter settings showed greater preference for vowel harmony over vowel disharmony, but the bias was only greater for variable conditions when there was a bias on μ. One possible

reason why the difference was greater for categorical learning than variable learning when the bias was on σ^2 creates different penalties for changing the weights. In a situation where there is a small amount of categorical data, having a bias that allows for larger changes may make it easier to show almost perfect learning with minimal training data.

The fact that a bias on μ best mirrors the human data suggests that learners come to the task with some prior experience with or preference towards the AGREE constraint. There is some research to support the idea that speakers have a general preference towards vowel harmony. Listeners tended to be faster and more accurate at identifying phonemes when they obeyed vowel harmony constraints (Kimper 2017), and English speaking infants tend to show preference for vowel harmony in speech segmentation tasks (Mintz et al. 2018). The present simulations support these results because all of the biased learning simulations show some preference for vowel harmony over disharmony, even in categorical conditions. This is different from many of the results of the artificial language learning studies comparing the learnability of a categorical vowel harmony to categorical vowel disharmony (Huang and Do 2023; Pycha et al. 2003; Skoruppa and Peperkamp 2011). However, in many cases in the simulations, the differences in learning for the harmony and disharmony conditions were quite small (e.g., 10%), just like the differences found in human learners under categorical conditions (Martin and Peperkamp 2020). For example, Pycha et al. (2003) showed a numerical advantage to vowel harmony over vowel disharmony, but this advantage was not statistically significant. Small biases may not always be detectable in an artificial language learning setting, which is subject to participant noise, and metacognitive judgments that may mask a small bias. Because participants in artificial language learning studies use a variety of strategies, some more implicit than others (Moreton and Pertsova 2024), it is reasonable to assume that a small bias for vowel harmony be more likely to be detectable under certain learning conditions. In the variable condition, participants receive evidence for both the harmony and the disharmony rules. If they already have some weight to the harmony-inducing constraint, this can push them to learn a harmony pattern better but make it even harder to learn a majority disharmony pattern.

The present study made us of the MaxEnt Harmonic Grammar Learning Tool to simulate learning biases for vowel harmony over vowel disharmony in variable vs. categorical exposure conditions. When the learner was given a small bias to weight vowel harmony before exposure to the training items, a bias emerged towards vowel harmony over vowel disharmony, and this bias was greater in variable conditions than categorical ones. These results may provide some insights into the nature of human learnability under natural and artificial settings.

References

BAKOVIĆ, ERIC. 2000. Harmony, dominance and control. PhD Dissertation Rutgers University.

DO, YOUNGAH.; and SHANNON MOONEY. 2022. Variation awaiting bias: Substantively biased learning of vowel harmony variation. *Journal of Child Language* 49. Cambridge University Press.397–407.

FINLEY, SARA. 2022. Generalization to novel consonants: Place versus voice. *Journal of Psycholinguistic Research* 51. Springer.1283–1309.

FINLEY, SARA. 2023. Modeling harmony biases in learning exceptions to vowel harmony. *Proceedings of the Linguistic Society of America* 8.5530–5530.

FINLEY, SARA. 2024. Vowel Harmony in Optimality Theory. https://academic.oup.com/edited-volume/58804/chapter/489198456.

GOLDWATER, SHARON.; and MARK JOHNSON. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm workshop on variation within Optimality Theory*, ed. by Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120.

HAYES, BRUCE.; and COLIN WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440. doi:10.1162/ling.2008.39.3.379.

HAYES, BRUCE.; COLIN WILSON.; and BENJAMIN GEORGE. 2009. Manual for Maxent grammar tool. *Online: https://linguistics. ucla. edu/people/hayes/MaxEntGrammarTool/ManualForMaxentGrammarTool. pdf.*

HUANG, TINGYU.; and YOUNGAH DO. 2023. Substantive bias and variation in the acquisition of vowel harmony. *Glossa: A Journal of General Linguistics*. doi:10.16995/glossa.9313. https://www.glossa-journal.org/article/id/9313/.

KIMPER, WENDELL A. 2017. Not crazy after all these years? Perceptual grounding for long-distance vowel harmony. *Laboratory Phonology* 8. doi:10.5334/labphon.47. https://www.journal-labphon.org/article/id/6206/.

MARTIN, ALEXANDER.; and SHARON PEPERKAMP. 2020. Phonetically natural rules benefit from a learning bias: a re-examination of vowel harmony and disharmony. *Phonology* 37. Cambridge University Press.65–90. doi:10.1017/S0952675720000044.

MARTIN, ALEXANDER.; and JAMES WHITE. 2021. Vowel harmony and disharmony are not equivalent in learning. *Linguistic Inquiry* 52.227–239. doi:10.1162/ling_a_00375.

MINTZ, TOBEN H.; RACHEL L. WALKER.; ASHLEE WELDAY.; and CELESTE KIDD. 2018. Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition* 171.95–107. doi:10.1016/j.cognition.2017.10.020.

MORETON, ELLIOTT.; and JOE PATER. 2012. Structure and substance in artificial-phonology learning, Part II: Substance. *Language and Linguistics Compass* 6.702–718. doi:10.1002/lnc3.366.

MORETON, ELLIOTT.; and KATYA PERTSOVA. 2024. Implicit and explicit processes in phonological concept learning. *Phonology*. Cambridge University Press.1–53. doi:10.1017/S0952675724000034.

OHALA, JOHN J. 1994. Towards a universal, phonetically-based, theory of vowel harmony. *Proceedings of the 3rd International Conference on Spoken Language Processing*, 491–494.

PYCHA, ANNE.; PAWEL NOWAK.; EURIE SHIN.; and RYAN SHOSTED. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. *West Coast Conference of Formal Linguistics (WCCFL)* 22.101–113.

SKORUPPA, KATRIN.; and SHARON PEPERKAMP. 2011. Adaptation to novel accents: Feature-based learning in context-sensitive phonological regularities. *Cognitive Science* 35.348–366. doi:https://doi.org/10.1111/j.1551-6709.2010.01152.x.

WALKER, RACHEL. 2012. Vowel Harmony in Optimality Theory. *Language and Linguistics Compass* 6.575–592. doi:10.1002/lnc3.340.

WHITE, JAMES CLIFFORD. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language* 93.1–36.

WILSON, COLIN. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30.945–982. doi:10.1207/s15516709cog0000_89.