December 2020

Emergent Typological Effects of Agent-Based Learning Models in Maximum Entropy Grammar

Coral Hughto

EMERGENT TYPOLOGICAL EFFECTS OF AGENT-BASED LEARNING MODELS IN MAXIMUM ENTROPY GRAMMAR

A Dissertation Presented

by

CORAL HUGHTO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2020

Linguistics

EMERGENT TYPOLOGICAL EFFECTS OF AGENT-BASED LEARNING MODELS IN MAXIMUM ENTROPY GRAMMAR

A Dissertation Presented

by

CORAL HUGHTO

Approved as to style and content by:

Gaja Jarosz, Co-chair

Joe Pater, Co-chair

John Kingston, Member

Brendan O'Connor, Member

Joe Pater, Department Chair
Linguistics

ACKNOWLEDGEMENTS

I am extremely grateful to everyone in my life for their love and support, in all of the ways it manifested. No words could possibly do justice to the depth of my love and appreciation, and I hope I can be forgiven for my brevity. Each word here carries the weight of six years of emotional and intellectual labor, and anyone not mentioned here is not forgotten in my heart. I could not have done this without any of you.

I would like first to thank my dissertation committee - Gaja Jarosz, Joe Pater, John Kingston, and Brendan O'Connor - without whom this certainly would not have been possible. Thank you all so much for your support and your feedback. Robert Staubs also deserves thanks here, as a member of my first GP committee and as a coauthor on earlier iterations of this work.

To Gaja and Joe especially I owe more than I can express. Joe saw me through this project from start to finish, and it would not have happened without him. Part of the reason I came to UMass over other options was because of my conversations with him during my visit as a prospective student. Gaja's guidance and support have been invaluable as I struggled to piece together this puzzle. I am so honored to have had her as a mentor. Thank you both for telling me I could, when I didn't believe it.

I would also like to thank the other faculty members in linguistics for their support in my time at UMass, especially including Jeremy Hartman, as the chair of my GP committees, John McCarthy, for his role as a mentor in my early years as a grad student, and to Kristine Yu, for her boundless enthusiasm and dedication to her students.

Tom Maxfield and Michelle McBride deserve special thanks for all of the administrative support they provide for the department. I would not have survived this

process without their guidance in navigating the university bureaucracy, and their dedication to helping the department run smoothly, and to helping the graduate students survive the program. Thank you both for being our advocates and allies.

The support and friendship of my cohort has also been a special part of my graduate career, and I am so proud of all eight of us! So much love to Caroline Andrews, Sakshi Bhatia, David Erschler, Ivy Hauser, Jyoti Iyer, Leland Kusmer, and Katia Vostrikova. I will miss our cohort get-togethers, and I wish everyone the best of luck in all of our future endeavors. Special thanks to Ivy for being the best officemate I could imagine, and for putting up with me for six years.

Many other (former and current) UMass linguistics graduate students have supported me as well, including Robert Staubs, Claire Moore-Cantwell, Presley Pizzo, Brandon Prickett, Andrew Lamont, Megan Somerday, Brian Smith, Aleksei Nazarov, and Amanda Rysling. Thanks also to those members of the Phonetics/Phonology Reading Group crew, the Spectrogram Reading Lunch crew, and the trivia night crew who I haven't yet mentioned: including Amanda Hayes, Erika Mayer, Maggie Baird, Carolyn Anderson, Arjun Guha, Leah Chapman, Chris Hammerly, Esra Yarar, Shay Huckleberry, Michael Wilson, Kaden Holladay, Katie Tetzloff, and Max Nelson, as well as to others that are not listed here, but are not forgotten. The friendships and support we gave each other will stay with me always. PRG and Spectrolunch reminded me that work could still be fun, and trivia nights reminded me that I could have fun outside of work.

Thanks also to my advisors and mentors from Indiana University: Stuart Davis, Dan Dinnsen, Judith Gierut, and Michelle Morrisette, without whom I would never have made it to graduate school in the first place. Thanks also to my fellow IU alumni Juliet Stanton and Emily Mange, and to my fellow graduate students Charlie O'Hara, Betsy Pillion, Caitlin Smith, Jesse Zymet, and anyone else who is not mentioned here. Your friendship and comradery has meant the world to me.

I would also like to thank the staff at Provisions, for keeping me supplied with wine and charcuterie, to the staff at the Lady Killigrew Cafe, where I spend a lot of time writing and grading, and to every other bar and coffee shop that let me work on their premises. I would also like to thank everyone at Sattva Center for Archery Training (formerly Amherst Archery Academy) for being a home outside of linguistics.

Finally, I also need to thank my family: my parents Tracy and Chris, and my sister Alyssa, for their love and support. I need to thank my cats Largo and Sinclair, for fluffy cuddles, and for every paragraph I had to reconstruct because they sat on my keyboard. And thanks especially to my wife, Io. We have built a wonderful life together, and I hope we have many years left to come.

ABSTRACT

EMERGENT TYPOLOGICAL EFFECTS OF AGENT-BASED LEARNING MODELS IN MAXIMUM ENTROPY GRAMMAR

SEPTEMBER 2020

CORAL HUGHTO

B.A., INDIANA UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Gaja Jarosz and Professor Joe Pater

This dissertation shows how a theory of grammatical representations and a theory of learning can be combined to generate gradient typological predictions in phonology, predicting not only which patterns are expected to exist, but also their relative frequencies: patterns which are learned more easily are predicted to be more typologically frequent than those which are more difficult.

In Chapter 1 I motivate and describe the specific implementation of this methodology in this dissertation. Maximum Entropy grammar (Goldwater & Johnson, 2003) is combined with two agent-based learning models, the iterated and the interactive learning model, each of which mimics a type of learning dynamic observed in natural language acquisition.

In Chapter 2 I illustrate how this system works using a simplified, abstract example typology, and show how the models generate a bias away from patterns which rely on cumulative constraint interaction ("gang effects"), and a bias away from variable patterns. Both of these biases match observed trends in natural language typology and psycholinguistic experiments.

Chapter 3 further explores the models' bias away from cumulative constraint interaction using an empirical test case: the typology of possible patterns of contrast between the fricatives /s/ and /ʃ/. This typology yields five possible patterns, the rarest of which is the result of a gang effect. The results of simulations performed with both models produce a bias against the gang effect pattern.

Chapter 4 further explores the models' bias away from variation using evidence from artificial grammar learning experiments, in which human participants show a bias away from variable patterns (e.g. Smith & Wonnacott, 2010). This test case was chosen additionally to disambiguate between variable behavior within a lexical item (variation), and variable behavior across lexical items (exceptionality). The results of simulations performed with both learning models are consistent with the observed bias away from variable patterns in humans.

The results of the iterated and interactive learning models presented in this dissertation provide support for the use of this methodology in investigating the typological predictions of linguistic theories of grammar and learning, as well as in addressing broader questions regarding the source of gradient typological trends, and whether certain properties of natural language must be innately specified, or might emerge through other means.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF TABLEAUX

Figure **Page**

CHAPTER 1
INTRODUCTION

1.1 Overview

In this dissertation, I demonstrate a methodology for generating gradient typological predictions through the combination of a theory of grammar with a theory of learning. The theory of grammar defines the set of possible linguistic representations, while the theory of learning introduces biases towards some of those possibilities and away from others. I show how this methodology can be used to investigate and account for observed cross-linguistic trends in phonological typology, focusing in particular on the distribution of gang effect patterns and patterns of variation. This approach crucially models typological trends as emerging from the interaction between grammar and learning, rather than needing to be coded explicitly into either, and has the flexibility to allow for testing the consequences of different theories of learning and representation.

In this first chapter, I begin by introducing the problem of probabilistic trends in linguistic typology. Much of the existing work on typology focuses on generating categorical predictions, distinguishing only between possible and impossible patterns, but this approach misses many interesting observations about probabilistic trends in typology: some linguistic patterns are more common than others. I next define the specific implementation of the methodology used throughout this dissertation to generate probabilistic typological predictions, first introducing the theory of grammar assumed here, Maximum Entropy grammar (MaxEnt; Goldwater & Johnson, 2003), as well as the two agent-based learning models, the iterated learning model (see e.g.

Kirby & Hurford, 2002; Kirby et al., 2014) and the interactive learning model (see e.g. Pater, 2012). I then detail how the combination of MaxEnt with these learning models is used to generate gradient typological predictions.

In the chapters that follow, I illustrate the application of this methodology using a number of test cases. Chapter 2 provides an analysis of how this model works using a simple, abstract typology, and details two types of bias produced by the model: a bias away from gang effect patterns, which rely on cumulative constraint interaction, and a bias away from variable patterns. Chapter 3 further investigates the bias away from gang effects, showing how this bias aligns with trends observed in natural language typology, focusing in particular on the case of the typology of constrast patterns between /s/ and /ʃ/. Chapter 4 further investigates the bias away from variable patterns, showing how this bias aligns with evidence from artificial language learning experiments. Chapter 5 concludes the dissertation, providing discussion of some implications of this work for framing the question of how grammar and learning interact, and how typological trends arise, as well as discussing the flexibility of this methodology and how it can be extended to additional assumptions and theories.

1.2 Background

One of the central goals of generative phonology is to accurately predict the range of attested natural languages. This has traditionally taken the form of a categorical typology, in which all of the patterns that can be represented by the grammatical theory are predicted to be possible, attested natural languages, and all those that cannot be represented are predicted to be impossible. In other words, a sufficiently explanatory theory of grammar should be able to represent *all* and *only* attested languages. This approach, however, misses the important observation that some attested patterns are cross-linguistically more frequent than others. In light of this fact, a growing body of literature has shifted the focus from a categorical distinction to a more gradient approach, which seeks to predict relative frequencies among attested languages (see Bane & Riggle, 2008; Culbertson et al., 2012; Pater, 2012; Staubs, 2014, among others). Unattested languages, under this approach, can include both patterns which are predicted to be impossible to represent, and ones which are possible to represent but predicted to be rare or unobserved due to the influence of learning or other grammar-external factors.

Much of the work which follows this gradient approach generates gradient typological predictions through combining a theory of grammar with a theory of learning. The theory of grammar determines which patterns are possible or impossible to represent, and the theory of learning introduces learning biases, either explicit or implicit, which make some possible patterns easier to learn than others. Patterns which are easier to learn are predicted to be more typologically common than patterns which are more difficult to learn, and thus would be less likely to be transmitted faithfully. This overall approach is general enough to be applied with any combination of grammatical theory and learning theory, in phonology or other areas. Methods used in previous work have included constraint-based theories, both with ranked constraints as in classic parallel Optimality Theory (OT, Prince & Smolensky, 1993/2004; see e.g. Bane & Riggle, 2008; Stanton, 2016) and with weighted constraints as in Maximum Entropy Grammar (MaxEnt, Goldwater & Johnson, 2003; see e.g. Pater, 2012; Staubs, 2014), as well as Bayesian modeling (see e.g. Griffiths & Kalish, 2007; Dediu, 2009; Culbertson et al., 2012).

Stanton (2016), for example, combines parallel OT with a convergent implementation of the Gradual Learning Algorithm in OT (Magri, 2012; see also Boersma, 1997; Boersma & Hayes, 2001), and applies this system to learning in a stress typology, seeking especially to understand the underattestation of the midpoint pathology, an unattested stress system in which the stressed syllable is located on a central word-internal syllable in some words, but not others. Stanton's learner shows a bias against the midpoint pathology, requiring a longer period of learning to converge on midpoint patterns, compared to other patterns. The difficulties faced by a learner attempting to acquire a midpoint pattern, Stanton argues, plausibly explains their absence from the observed typology without needing to modify the grammatical theory to exclude these patterns, which would have unintended consquences for other, attested patterns.

Another example, Culbertson et al. (2012), conduct an artificial language learning experiment to test for learning biases in learning word order patterns, specifically in the ordering of Numerals and Adjectives relative to Nouns in a phrase. In the observed typology, consistent word orders in which the Adjective and Numeral have the same ordering relative to the Noun (Noun-Adj/Noun-Num and Adj-Noun/Num-Noun orders) are more common relative to inconsistent word orders in which the preferred positions of the Adjective and Numeral are different (Noun-Adj/Num-Noun and Adj-Noun/Noun-Num orders), and within the different-preference orders, the Adj-Noun/Noun-Num ordering is very rare compared to the Noun-Adj/Num-Noun order. The human participants in their experiment showed learning biases consistent with the typological observations, preferring consistent word orders to inconsistent word orders, and preferring the more common Noun-Adj/Num-Noun order to the rare Adj-Noun/Noun-Num order. Culbertson et al. support these findings with an explicit Bayesian learning model trained on the same data patterns as their experimental participants (see also Culbertson & Smolensky, 2012), showing how the learning biases that they propose explain the behavior of the human participants in the experiment, also work to produce the same behavior in the Bayesian learning model.

Although learning biases provide a rich source of explanation for typological trends, as evidenced by the work described above, Rafferty et al. (2013) argue against using learnability biases as the sole predictor for the greater frequency of a given pattern. The authors discuss two cases in which greater learnability of a pattern may not result in the greater predicted frequency of that pattern. In the first case, they argue that a more learnable pattern or concept which is unlikely to be generated spontaneously as the result of mislearning may not necessarily be frequently observed, as the greater likelihood of learning that pattern faithfully would be offset by the smaller chance of recovering that pattern once lost. They test this hypothesis in a memory test experiment, in which participants were asked to memorize a list of ten items, complete a short read and respondse task as a distractor, and then repeat their memorized list, giving their best guess on any items they couldn't remember so that their generated list also contained ten items. Participants were arranged into chains, so that the list generated by one participant was given to the next participant in the chain to memorize, allowing the experimenters to track the changes in the list as it was passed through the chain. There were four such chains, each consisting of 50 people. The list given to the first participant in each chain contained ten items: nine items that one would commonly buy at a grocery store (toilet paper, cheese, tomatoes, eggs, milk, lettuce, orange juice, bread, bananas), plus one distinctive, unrelated item (elephants). The results of this experiment found that while the distinctive item *elephants* was correctly remembered 95% of the time, while the grocery items were correctly remembered only 87% of the time, *elephants* eventually disappeared from each participant chain, and was never spontaneously generated by a later participant once lost. In contrast, the less commonly remembered grocery items were also more commonly spontaneously generated, and so had a chance of reappearing later in a chain if lost. Rafferty et al. (2013) thus assert that in addition to analyzing how likely a pattern is to be learned faithfully, researchers must analyze what happens when a pattern is not learned faithfully, which requires simulating the process of pattern transmission across a population.

In the second case, Rafferty et al. (2013) argue that a more learnable pattern or concept may not necessarily attain greater prevalence if there are a large number of alternative hypotheses in competition, as the combined probability of the alternatives could offset the effect of greater learnability. As support for this hypothesis, they consider the case of vowel harmony, a fairly common linguistic pattern where vowels in a word must share some phonological property. Some previous studies have demonstrated that humans show a learning bias favoring vowel harmony patterns over alternatives (e.g Finley & Badecker, 2009; Moreton, 2008), which leads Rafferty et al. to the question: why is vowel harmony not a lingusitic universal? The reason, they argue, lies in the fact that the constraints imposed by vowel harmony patterns greatly restrict the space of possible words, such that there are going to be many more possible languages which do not feature vowel harmony than do feature vowel harmony. Using a mathematical analysis, Rafferty et al. (2013) argue that a very strong bias towards vowel harmony would be needed in order for vowel harmony to become a linguistic universal in the face of so many competing non-harmonic patterns, and based on the findings from previous artificial language learning experiments as well as their own, they argue that the bias towards vowel harmony shown by human participants is not strong enough to make vowel harmony a dominant pattern type. Rafferty et al. (2013) thus assert that the existence of a learning bias towards a particular linguistic property does not necessarily mean that property will become a linguistic universal, or even a dominant property, depending on the strength of the learning bias and the number of competing possible hypotheses.

Overall, Rafferty et al. (2013) caution against assuming a straightforward connection between learning biases shown and typological trends, and describe the potential benefits of actually simulating the transmission of patterns across individuals, either in experimental contexts or with computational modeling, in order to gain a better understanding of how the learning biases for or against particular properties relate to the distribution of those properties across the typology. The work presented in this dissertation adopts this simulation-based approach to linking learning biases to typological trends, building on a number of previous studies that have investigated the consequences of modeling the transmission of language patterns across individuals through the use of agent-based learning models, in which learning is modeled as the exchange of data between simulated entities (see Zuraw, 2003; Wedel, 2011, for overviews of agent-based learning in linguistics). In particular, many of these studies model language evolution across generations, using the iterated learning model (Kirby & Hurford, 2002; Kirby et al., 2014, see §1.4 for more detail), in which a learner agent first learns from data provided by an immutable teacher agent, then becomes the teacher agent for a new learner agent, repeating this process across a chain of agents (see Dediu, 2009; Griffiths & Kalish, 2007; Rafferty et al., 2013; Staubs, 2014; Kirby et al., 2015, among others). Another approach, which I call in this dissertation the interactive learning model, simulates the interaction between agents exchanging data, each learning from the other's outputs (e.g. de Boer, 2000; Pater, 2012, see §1.4 for more detail).

The work presented in this dissertation builds in particular on the work by Pater (2012) and Staubs (2014), who use as a grammatical base MaxEnt, a constraint-based theory in which constraints are weighted and a probability distribution is defined over competing output candidates (see §1.3.1 for more formal details), combined with agent-based learning models. Pater (2012) combines MaxEnt with an implementation of the interactive learning model, using computational modeling to apply this approach to the problem of learning phrase headedness. The training data consisted of 4 phrase types, each of which could be right-headed or left-headed. The constraints used to represent this data in the MaxEnt grammars include a combination of general constraints which favor all phrases to be right-headed or left-headed, and phrase-specific constraints which prefer particular phrase types to be right-headed or left-headed. The results demonstrate that this learning model produces a bias towards regularization, such that a general pattern, in which phrases are either all right-headed or all left-headed, is preferred over a pattern with exceptions, in which different phrases can have different headedness preferences. This regularization bias, Pater argues, is stronger than would be expected without the influence of the learning model, and matches the observed cross-linguistic tendency towards consistent headedness across phrase types. The model results also show a bias towards deterministic grammars, where, for a given phrase type, the probability of producing each possible headedness order is pushed towards either 1 or 0.

Staubs (2014) combines the MaxEnt grammatical framework with an implementation of the iterated learning model. Among other patterns, Staubs applies the iterated learning model to the problem of learning stress window systems, in which main stress is required to fall on a heavy syllable, but only if it lies within some number of syllables away from the word edge; otherwise, stress defaults to the word edge. In the observed typology of stress window systems, two-syllable windows are more common than three-syllable windows, and larger window sizes are unattested. In weighted constraint grammars such as MaxEnt, stress windows can be represented as gang effects, which are patterns where multiple violations of lower-weighted constraints cumulatively outweigh a single violation of a higher-weighted constraint. Because of this property of cumulative constraint interaction, weighted-constraint grammars have been criticized for overpredicting the range of typological possibilities, relative to ranked-constraint theories. While ranked-constraint grammars cannot represent stress windows without special additional constraints, weighted-constraint grammars go to the other extreme, and can represent stress windows of any arbitrary length, which does not match the observed typology. Under an approach which assumes that the grammatical theory should be able to represent *all* and *only* possible patterns, the trade-off must be resolved either in favor of ranked-constraints, which can represent only but not all possible patterns, or weighted-constraints, which can represnt all but not only possible patterns. In combining MaxEnt grammar with the iterated learning model, however, Staubs (2014) found that the influence of learning produced a gradient bias against the gang effect interaction that results in stress windows, such that two-syllable stress windows had a higher predicted probability than three-syllable windows, and four-syllable windows and larger had a predicted probability at or approaching zero. These results reflect the observed gradient typology of stress windows, showing that appropriately restrictive predictions can be derived when grammar-external factors such as learning are taken into account, without needing to sacrifice representational power in the grammar.

The research presented in this dissertation builds on this previous work, and makes a more detailed investigation of the properties of combining the MaxEnt grammatical framework with the interactive and iterated agent-based learning models, and examines their predictions for gradient phonological typology. As detailed in Chapter 2 using a simple test typology, this work finds that both the iterated and interactive learning models produce similar biases in the predicted typologies: (1) a bias away from gang effect patterns, which rely on cumulative constraint interactions, where violations of a lower-weighted constraint can "gang up" on a violation of a higher-weighted constraint (explored further in Chapter 3), and (2) a bias towards more deterministic grammars, in which the probability of a given output candidate is pushed towards 1 or 0 (explored further in Chapter 4). The combined effect of these biases results in behavior that brings the predictions of these models closer to those of classic ranked-constraint OT grammars, in which only one candidate is chosen as optimal and constraint violations are not cumulative. Thus, the use of a weighted-constraint grammar such as MaxEnt for the grammatical base allows for the representation of gang effect patterns as with the stress windows case in Staubs (2014), as well as variation and exceptionality as with the phrase headedness case in Pater (2012), while the addition of the agent-based learning models reduces the overprediction problem by introducing learning biases against gang effects and variation, bringing the typological predictions more in line with observed empirical trends.

The remainder of this chapter will proceed as follows. In §1.3, I discuss the details of the MaxEnt grammatical framework and its similarities to and differences from categorical Harmonic Grammar (see e.g. Legendre et al., 2006; Pater, 2009). In §1.4 I describe the iterated and interactive learning models, how they are implemented in the dissertation (§1.4.1 and §1.4.2 respectively), and how probabilistic typological predictions are generated with this system (§1.5).

1.3 Typology and Maximum Entropy grammar

In categorical constraint-based grammars, the predicted typology of a given system is typically held to be the set of all unique sets of output candidates which can be simultaneously selected as optimal under some permutation of the constraints, whether they are ranked, as in classic parallel OT, or weighted, as in Harmonic Grammar. This approach has been sufficient under traditional assumptions and methods, in which the primary aim of typology is to account for possible and impossible categorical phonological patterns. However, not all natural language phenomena are categorical: languages can exhibit systematic patterns of variation and exceptionality, language input to new learners is noisy, and, as I will discuss further in this work, not all possible language patterns are observed with equal frequency. Probabilistic grammatical models such as Maximum Entropy grammar are well equipped to analyze such non-categorical phenomena; however, the complications that arise when trying to predict a typology over probabilistic patterns have raised many methodological questions, and have prompted a more thorough investigation of the representational capabilities of probabilistic grammatical models, and how their typological predictions relate to the more well-understood categorical grammatical models (see e.g. Pater, 2016; Anttila & Magri, 2018). In §1.3.1, I will first define the MaxEnt grammatical framework assumed in this work, and then in §1.3.2 I will discuss some of what is already known about the representational power of MaxEnt, and how it compares to a related categorical model, HG.

1.3.1 Maximum Entropy grammar

The grammatical framework assumed in this work is Maximum Entropy grammar (MaxEnt; Goldwater & Johnson, 2003). MaxEnt is a probabilistic version of Harmonic Grammar[1] (HG; see e.g. Legendre et al., 2006; Pater, 2009), in which the constraints are assigned numeric weights, rather than ranked as in classic Optimality Theory (OT; Prince & Smolensky, 1993/2004). In both MaxEnt and HG, each output candidate receives a Harmony (H) score, which is defined as the weighted sum over that candidate's constraint violations, where each violation is multiplied by the constraint weight. Constraint violations are restricted to be negative integers, and constraint weights are restricted to be positive real numbers. Harmony is calculated as in the equation in (1.1), where K is the set of constraints, W_k is the weight of constraint k and $V_k(x)$ is the number of violations of constraint k incurred by candidate x. This work adopts a standard set of assumptions in which constraint weights are restricted to be non-negative, and constraint violations are negative integers (see Prince, 2003; Boersma & Pater, 2016; Pater, 2009; Potts et al., 2010, for discussion).

$$H(x) = \sum_{k \in K} W_k \cdot V_k(x) \tag{1.1}$$

In HG, the optimal candidate is the one with the highest (closest to zero) H score. Because constraint weights are postive and violations are negative, the maximum H score is zero, and so the optimal candidate is the one which incurs the lowest cumulative penalty. MaxEnt, on the other hand, does not define an optimal candidate;

[1] Here and throughout the dissertation I use the term HG to refer specifically to weighted-constraint grammars which define Harmony scores over output candidates in the manner described in this section.

rather, the H scores are used to calculate a probability distribution over output candidates. An output candidate's probability is defined as the proportion of the exponential of its H score, out of the total sum over the exponentiated H scores of it and its competitors. This is calculated as in the equation in (1.2), where $X = \{x_1...x_n\}$ is the set of competing candidates. Note that in MaxEnt, probabilities can approach, but never exactly equal 1 or 0.

$$p(x) = \frac{exp(H(x))}{\sum_{i=1}^{n} exp(H(x_i))} \tag{1.2}$$

One advantage of weighted constraints over ranked constraints is that weighted constraints are compatible with the use of existing convergent learning algorithms for finding a set of constraint weights consistent with a given set of data (see e.g. Jäger, 2007)). Learning algorithms exist for ranked-constraint systems (e.g. Pulleyblank & Turkel, 1996; Prince & Tesar, 1999; Tesar & Smolensky, 2000), but none of these have both the properties of being able to handle noisy data, and being proven to converge (see Jarosz, 2016b,a, for overviews).

The learning algorithm used in this work is a gradual, error-driven learning algorithm related to Stochastic Gradient Ascent (see e.g. Jäger, 2007). Learning data take the form of input-output pairs which are sampled from the probability distribution defined by the MaxEnt grammar. For a given input, if the learner's sampled input does not match the target's sampled input, then the learner agent updates its constraint weights via the update rule given in (1.3), where $W_k^{(t)}$ is the weight of constraint k at the current timestep and $W_k^{(t-1)}$ is the weight of constraint k at the previous timestep, x_T is the input-output pair generated by the target, x_L is the input-output pair generated by the learner, and η is the learning rate. The updated constraint weights are calculated by subtracting the violation profile of the learner's erroneous form from the violation profile of the teacher's form, scaling the difference by a learning rate, and then adding it to the learner's old constraint weights. This update serves to raise the weights of constraints which favor the teacher's form, and lower the weights of constraints which favor the learner's erroneous form. By consequence, this also places higher probability on the teacher's form, and lower probability on the learner's erroneous form. This update rule is related to the Delta Rule in connectionist modeling (Rumelhart & McClelland, 1986) and the Perceptron update rule in natural language processing (see e.g. Johnson, 2007), and is very similar to the Gradual Learning Algorithm for Stochastic OT (Boersma, 1997; see also Pater, 2008; Boersma & Pater, 2016) and to Stochastic Gradient Ascent (Jäger, 2007).

$$W_k^{(t)} = W_k^{(t-1)} + \eta(V_k(x_T) - V_k(x_L)) \tag{1.3}$$

One criticism of weighted-constraint systems is that they allow for cumulative constraint interactions, or "gang effects", which has been argued to be an undesirable feature which predicts unattested patterns (Legendre et al., 2006, but also see Pater, 2009). In a gang effect, multiple violations of one or more lower-weighted constraints can cumulatively outweigh a violation of a higher-weighted constraint. This type of interaction is not possible with ranked constraints, as a candidate which incurs a fatal violation of a high-ranked constraint is ruled out, no matter how many violations of lower-ranked constraints are incurred by its competitors. However, there are some attested patterns that can be represented as gang effects without needing to assume additional special constraints, for example stress windows (see e.g. Staubs, 2014) and the Elsewhere Neutralization pattern discussed in Chapter 3 (see also Carroll, 2012). While it would be possible to represent these patterns in ranked-constraint grammars with the addition of special constraints to handle these cases, analyzing them as gang effects negates the need to handle each case individually - as they arise from interactions between more widely motivated constraints - with the tradeoff being that these patterns appear to arise much less commonly than might be expected. It would seem desirable, then, to build a model which allows for the existence of gang effects, but predicts them to be much less common than other pattern types. As I show in my work, this can be done by combining MaxEnt with the iterated and/or interactive learning models - these models show emergent biases which reduce the predicted probability of gang effect patterns.

In the next section, I turn to the question of the representational power of MaxEnt, and what is known about how the typological properties of MaxEnt compare to categorical HG.

1.3.2 MaxEnt representational capabilities

In order to accurately evaluate the influence that learning will have on the typological predictions of a given system, it is important to first understand the typological properties of the base grammatical model by itself; here, MaxEnt grammars. With categorical constraint-based grammatical models, such as classic parallel OT and HG, it is possible to enumerate the set of possible patterns by defining the set of all sets of unique output forms which can simultaneously be chosen as optimal under some permutation of the constraint rankings or weightings. In probabilistic grammatical models like MaxEnt, this enumeration is impossible implement because MaxEnt does not define a set of optimal candidates, but rather defines probability distributions over sets of competing output candidates. The most obvious analogue to applying this method in MaxEnt would be to define the set of all sets of unique probability distributions over competing output candidates, however, this is computationally intractable, because probabilistic grammatical models can produce an infinite number of possible probability distributions.

In order to avoid these difficulties, any investigation into the representational properties of MaxEnt must take a different approach to delimiting the set of possible and impossible patterns. Because MaxEnt is a probabilistic variation of HG, some similarities and differences between MaxEnt and HG can be drawn. As with all constraint-based grammars, the representational power of MaxEnt is restricted by the content of the constraint set and the nature of constraint interaction. Beyond this, the predictions of MaxEnt converge and diverge from categorical HG in various ways.

One similarity between MaxEnt and HG is that, as a probabilistic extension of categorical HG, MaxEnt shares the same predicted set of categorical patterns. In a probabilistic grammar model, categorical patterns arise in the special case when the probability of each candidate approaches one or zero. Because probability in MaxEnt is calculated as a proportion of exponentiated harmony, out of the sum over the exponentiated H scores of all competing candidates, the output candidate which receives the highest probability among its set of competitors must be the one with the highest Harmony score. Consequently, the set of output candidates which can receive probability approaching $p = 1$ in MaxEnt are all and only those which HG can make optimal under some constraint weighting.

One primary difference between MaxEnt and HG is that MaxEnt defines a probability distribution over competing output candidates, where HG defines an optimal output. One consequence of MaxEnt's probability calculation is that, since all competing candidates make up some proportion of the total Harmony, MaxEnt can assign non-zero probability to harmonically-bounded candidates. However, just as these candidates cannot be made optimal in categorical HG because their H scores will always be at least as great as their competitors, MaxEnt cannot assign probability to a harmonically bounded candidate which is greater than the probability of a non-harmonically bounded competing candidate. It is not obvious that assigning probability to harmonically-bounded candidates is undesirable, as it allows for accounts of local optionality and speech errors (see Pater, 2016, for discussion). Similarly, among a set of competing output candidates, MaxEnt will always assign higher probability to candidates with higher H scores. In cases of variation, then, the prediction is that output candidates will vary with each other in proportion to how well-formed they are under that constraint weighting.

1.3.3 Defining typology and evaluation metrics

As discussed earlier, in categorical models such as HG, the generated typological predictions consist of the set of all possible sets of output candidates which can be simultaneously optimal. In MaxEnt, however, generating such a typology is intractable because, rather than defining an optimal candidate, MaxEnt defines a probability distribution over competing output candidates. Even after defining some similarities and differences between HG and MaxEnt, the question still remains of how to operationalize a definition of "typology" in MaxEnt which can be practically implemented in order to explore its predictions. For the purposes of this dissertation, I define "typology" as a combination of two metrics, one which calculates a prediction over pattern types, and one which calculates a prediction over probability distributions.

The **pattern type** metric is defined as the set of all sets of output candidates which can simultaneously receive the highest probability among their set of competitors. This is very similar to the way in which typologies are calculated in OT or HG, which define typology as the set of all possible combinations of output candidates which can be simulataneously optimal. In fact, as discussed earlier, the categorical patterns predicted by MaxEnt are the same as those predicted in HG. However, for MaxEnt we are abstracting away from the probability distribution in ignoring the degree to which the highest probability candidate is preferred over its competitors. This abstraction makes it possible both to enumerate the set of pattern types that can be represented in MaxEnt, as well as to compare that set to the predictions made by other grammatical models.

Although the abstraction away from probabilities is useful for the pattern type metric, any investigation into the typological predictions of MaxEnt is incomplete without a consideration of the probabilistic component, and its interactions with the set of predicted pattern types. For this purpose, I use one of two measures: the average probability assigned to the highest probability output candidates for each input, or conditional entropy, which is a measure of how evenly the probability mass is distributed among competing output candidates, across items in the grammar. Both of these metrics define a value for a given grammatical state; the typology is then a distribution over possible values for these metrics.

Chapters 2 and 3 use the **average highest probability** measure, calculated as in the equation in (1.4), where N is the set of input forms in the lexicon, $len(N)$ is the number of input forms in the lexicon, and $X = GEN(n)$ is the set of competing output forms for a given input form n. This equation takes the average over the probabilities of the highest probability output form for each input. Average highest probability values approaching one indicate a deterministic grammar in which probability is concentrated on one output candidate in each set of competitors. Because the lowest possible probability that can be assigned to the highest probability output candidate is in the case where probability is distributed equally among the output candidates, the lower bound for this value is determined by the number of competing output candidates for each input.

$$\frac{1}{len(N)} \sum_{n \in N} \max_{x \in X} p(x|n) \tag{1.4}$$

Chapter 4 uses the **conditional entropy** measure, which is calculated as in the equation in (1.5), where N is the set of input forms in the lexicon, and $X = GEN(n)$ is the set of competing output forms for a given input form n. This equation is adapted from Smith & Wonnacott (2010), where entropy is measured in bits, and so uses a base 2 logarithm. Conditional entropy calculates the sum over the log conditional probabilities of each competing output, given an input, weighted by the joint probability of the input-output pair. For the purposes of this work, I make the simplifying assumption that the probability of each input form is $1/len(N)$, where $len(N)$ is the number of input forms. In natural language, however, word frequencies across the lexicon typically follow a Zipfian distribution (Zipf, 1949). Note that because in MaxEnt the output forms are conditioned on the input forms, the input probabilities and output probabilities are not independent. The lower bound on conditional entropy values is zero[2], which indicates a fully deterministic system in which probability is concentrated on one output candidate among the set of competitors, while greater and greater values indicate systems which are less and less predictable. In the special case where there are only two possible outputs for each input, which is the case for all of the examples used throughout this dissertation, the upper bound on conditional entropy is one, which indicates a system where every output has equal probability.

$$H(x|n) = -\sum_{n \in N} \sum_{x \in X} p(n,x) log_2 p(x|n) \tag{1.5}$$

Table 1.1 shows some example probability distributions, and their associated values according to the average highest probability measure and the conditional entropy measure. In these examples, there are two inputs, I_1 and I_2, and there are two competing output candidates for each input, A and B. Example 1 shows a case where probability is distributed equally within each set of competing output candidates, yielding an average highest probability of 0.5, and a conditional entropy value of 1. Example 2 shows a case at the other extreme, where both inputs place a probability approaching 1 on candidate A, and a probability approaching 0 on candidate B, yielding an average highest probability approaching 1 and a conditional entropy value approaching 0. These values are rounded for readability, but note that because $log_2(0)$ is undefined, conditional entropy in the event of a probability of zero is treated as

[2] Conventionally, the expression $(0)log_2(0)$ is treated as being equal to 0.

being zero. However, in MaxEnt, probabilities can approach zero or one, but never exactly equal zero or one, so this absolute lower bound is unattainable.

Comparing Example 2 and Example 3 shows that neither of these measures is sensitive to which output candidates an input prefers, only to the probability values. In Example 2, input I_2 prefers candidate A, while in Example 3 I_2 prefers B; however, because the probability values remain the same, so do the values of the average highest probability and conditional entropy measures.

Table 1.1. Example probability distributions illustrating the average highest probability measure and conditional entropy measure

		Probability				
Input	Output	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5
I_1	A	0.5	1.0	1.0	0.75	1.0
I_1	B	0.5	0.0	0.0	0.25	0.0
I_2	A	0.5	1.0	0.0	0.75	0.5
I_2	B	0.5	0.0	1.0	0.25	0.5
Avg. Highest Prob.		0.5	1	1	0.75	0.75
Conditional Entropy		1	0	0	0.81	0.5

Example 4 in Table 1.1 shows a less deterministic case where both inputs prefer candidate A with a probability of 0.75, yielding an average highest probability of 0.75 and a conditional entropy value of 0.81. Comparing Example 4 and Example 5 shows that the conditional entropy measure is sensitive to the degree to which the system overall is deterministic, while the average highest probability measure is not. In Example 5, input I_1 deterministically prefers candidate A with a probability of 1.0, but input I_2 places equal probability on A and B. The distribution in Example 5 yields the same average highest probability as in Example 4 (0.75), but a lower conditional entropy value (0.5), because of the equal distribution of probability over the output candidates for input I_2.

The combination of the pattern type and probability distribution metrics yields the predicted typology, where the pattern type metric defines the set of possible pattern types, and the probability distribution metric defines the range of possible associated highest average probability or conditional entropy values. This dissertation is concerned with defining the typological predictions generated by MaxEnt, and with exploring how the introduction of learning biases can influence those predictions. The next section introduces the learning models used throughout this work - the Iterated learning model and the interactive learning model - and discusses the details of their implementation.

1.4 Learning models

Having now established the properties of the representational system used in this dissertation, Maximum Entropy Grammar, I now turn towards defining the other half of the system used to generate gradient typological predictions - the learning models. Language in humans is fundamentally a social phenomenon, which must be learned from other humans. Over the course of language acquisition, learners will interact with other members of their social network who will themselves display varying levels of competence with the linguistic conventions of the ambient community. In choosing learning models to use in this dissertation, I rely on a substantial body of previous research on agent-based modeling, which models the aggregated population-level or group-level properties of a network of agents as emerging, sometimes non-transparently, from the properties of those individual agents as they evolve and interact with their environment. Agent-based modeling is a general framework that has been used across a wide variety of fields, including biology (e.g. An et al., 2009; Mansury et al., 2006; Walker et al., 2004, among many others), economics (e.g. Tesfatsion, 2002; Buchanan, 2009; Arthur, 2015, among many others), and the social sciences more broadly (e.g. Bankes, 2002; Epstein & Axtell, 1996; Bonabeau, 2002, among many others). The basic structure of an agent-based model defines the knowledge held by each agent and the rules for how agents interact with each other and their environment, then studies the behavior that emerges in the system given those parameters. In this dissertation, I implement two kinds of agent-based learning models that have been previously used in linguistics-related research: the iterated learning model and the interactive learning model. In the rest of this section, I discuss previous work with each of these models, and discuss the types of interaction dynamics each of them is designed to model, in the context of language learning and evolution.

The first of these, the iterated learning model (see Kirby & Hurford, 2002; Kirby et al., 2014, for overviews), is a model of the vertical transmission of a language from an experienced user of that language to an inexperienced user of that language. Once the inexperienced user becomes an experienced user, they in turn transmit their language to a new inexperienced user, and the process repeats in a cycle. In natural language acquisition, this is analogous to the passing of a language through generations, as parents teach their children, who may in turn grow up to become parents themselves. One of the earlier uses of this type of model comes from Hare & Elman (1995), who used connectionist models to investigate potential causes of the evolution of English verb inflection from the more complex paradigms in Old English to the modern English system, which has one regular inflection (e.g. *play/played*) and several classes of irregulars (e.g. *sing/sang*, *take/took*). They hypothesized that lexical frequency and phonological cues to class membership interacted with phonological changes in English, causing certain words to be reanalyzed as members of other classes, and certain classes to be collapsed. In a technique they called "generational" learning, they trained connectionist networks to generate inflected forms of a given verb, and chained the networks together by using the output of one network, after learning, as the input to the next network. Thus, the changes and/or errors in the system learned by one network get transmitted to the next network in the chain. The evolution of the inflection classes learned by the networks across these chains was consistent with the observed historical developments, supporting their hypothesis for the triggers of the diachronic shift.

More recently, Kirby (2001) introduced the iterated learning model, under that name, as a model of language evolution, with the aim of investigating whether certain design features of human language, such as semantic compositionality and regularity in morphosyntax, might have emerged as advantageous adaptations, rather than being innately-specified constraints on language. Kirby (2001) detail the results of iterated learning simulations in which learning agents are arranged in chains, such that the system learned by one agent is used to generate training data for the next agent in the chain. The data space consists of a set of two components, a and b, which can each take a range of five values $1-5$, and when combined yield 25 possible meanings (e.g. a_1b_1, a_1b_2, etc.). Each meaning has an associated label, and each agent is trained on 50 label-meaning pairs randomly sampled by the previous agent in the chain. Given the random sampling of the training data, each learning agent will only observe labels for a subset of the possible meaning space, but will need to produce labels for all possible meanings. For the initial agent in a chain, the labels are randomly generated strings of characters, and so there is no compositional structure to aid the agent in assigning labels to the unobserved meanings. As this initial, unstructured system is transmitted across agents in the chain, the labels generated by the agents gradually develop morpheme-like internal structure such that, for example, the labels for all meanings containing the component a_1 contain the same substring. Thus, compositionality emerges as an adaptation to the learnability pressures that arise from having to learn a fully expressive system on the basis of only a subset of all possible meanings.

The iterated learning model has since been applied in other computational modeling research on topics such as category structure learning (Griffiths et al., 2008), biases in the typology of stress (Staubs, 2014), variation and language change (e.g.

Reali & Griffiths, 2009; Sonderegger & Niyogi, 2013; Smith et al., 2017; Ferdinand et al., 2019), as well as in various works exploring the parameters and assumptions of the iterated model, including the effects of varying the number of agents and, with Bayesian agents, how the learner selects a hypothesis (e.g. Griffiths & Kalish, 2007; Dediu, 2009; Smith, 2009; Burkett & Griffiths, 2010; Rafferty et al., 2013; Perfors & Navarro, 2014). There are a number of studies, however, which assert that the vertical transmission chains characteristic of the iterated learning model present an incomplete picture of language change and evolution, because they abstract away from the potentially crucial effects of communication and interaction between agents during learning (Dediu, 2009; Kirby et al., 2015; Smith et al., 2017). Kirby et al. (2015), for example, show how the emergence of semantic compositionality from an initially random assignment of labels to meanings, as described in Kirby (2001), depends not only on the learnability pressure exerted by language transmission in iterated learning, favoring simpler systems with fewer labels to learn, but also on the pressures for expressibility in communication that arise as agents need to convey a specific meaning, favoring expressive systems which can uniquely identify all meanings. Compositionality is the compromise, with a small set of labels which can be freely recombined to assign unique signifiers to all meanings.

The second model explored in this dissertation addresses this dynamic of communication and interaction between agents, and I refer to it throughout this work as the interactive learning model. This interactive model simulates parallel, reciprocal interactions between language users. Each agent is both learning from data produced by other agents, and producing data for other agents to learn from. In natural language, this dynamic has analogues in linguistic accommodation, and in the spread of new words and linguistic conventions. As described earlier in this chapter, Pater (2012) applied an interactive learning approach to modeling the emergence of a preference for consistent right- or left-headedness in syntactic phrases, which correlates with the observed tendencies in natural language typology. One of the earlier uses of this type of model in linguistics is de Boer (2000), who used an interactive, agent-based model to investigate the possible origins of the systematic structural tendencies observed in sound inventories cross-linguistically. de Boer (2000) focused on vowel inventories, which display a number of cross-linguistic tendencies over the frequencies of particular vowels (e.g. [a] is much more common than [œ]) and over structural symmetries (e.g. an inventory containing the mid back lax vowel [ɔ] is much more likely to also contain the front counterpart [ɛ] than not). Some have argued that these systematic tendencies reflect innate restrictions on possible linguistic structures, however, de Boer (2000) shows how these tendencies can emerge through the interactions between agents in a population, without needing to posit an a priori restriction.

In the model implemented in that paper, the agents represent vowel categories as prototypes in a simplified acoustic space, and interact by playing imitation games. In each round, one agent, the "initiator" chooses a vowel prototype from its inventory, and synthesizes a set of acoustic values for that vowel, with added noise. The other agent, the "imitator", maps the heard acoustic values to the nearest prototype in its own inventory, then synthesizes a set of acoustic values for the perceived prototype. The initiator then maps the acoustic values heard from the imitator to the nearest prototype in its inventory. If the perceived prototype matches the initially produced prototype, the imitation game was successful, and the imitator shifts that vowel prototype to be closer to the acoustic values produced by the initiator. If the imitation game was unsuccessful, the imitator either shifts the unsuccessful prototype, or forms a new prototype, depending on how successful that category had been used in previous games. As de Boer (2000) shows, the sets of vowel prototypes learned by the interacting agents in these simulations show systematic structural tendencies similar to those observed in natural language vowel inventories.

Some more recent work, Zuidema & de Boer (2009) and de Boer & Zuidema (2010), applies an interactive learning approach to another feature of language which has been argued to stem from an innate constraint on linguistic structure. The authors investigate the possible origins of combinatorial structure in phonology, referring to the use in language of small, meaningless units (e.g. phonemes) which can be recombined to form larger, meaningful units (e.g. words). In the models implemented in these papers, the agents in the interacting populations encode signals as trajectories (analogous to the word level) between a start point and end point (analogous to the phoneme level) in an abstract two-dimensional plane. In this type of space, a system with no combinatorial structure would take the form of a set of trajectories which each have a unique start and end point; that is, each trajectory must be encoded uniquely and holistically (e.g. the set $\{AA, BB, CC\}$, where the first element in each pair is the coordinates of the start point and the second element is the coordinates of the end point). A system with combinatorial structure, on the other hand, would consist of a unique set of trajectories connecting a reusable set of start and end points (e.g. $\{AA, BB, AB\}$). Agents are initialized with a set of holistic, non-combinatorial trajectories, and then interact by playing imitation games of the type described in the discussion of de Boer (2000). As the agents interact, the start and end points of the set of trajectories begin to cluster, while each individual trajectory remains unique, showing an emergence of combinatorial structure.

Reali et al. (2014) also take an interactive learning approach, and investigate the seemingly conflicting observations that languages with larger numbers of speakers tend to be structurally simpler, but yet have a greater number of content words, than languages with smaller numbers of speakers, which tend to display more complex linguistic structures, but also have a smaller vocabulary (see e.g. Lupyan & Dale, 2010; Dale & Lupyan, 2012; Dryer & Haspelmath, 2013; Trudgill, 2011; Wray & Grace, 2007; MacWhorter, 2002; Pawley, 2006). They hypothesize that complex linguistic structures require more exposure to learn, and thus are more likely to spread successfully across a smaller group of speakers, who might interact with the same individuals frequently, than in a larger group of speakers, who might interact with a large number of individuals, but infrequently with a given individual. In other words, a smaller network of speakers affords more exposure to a given innovation, making it more likely to spread to new speakers. Lexical items, on the other hand, are more readily innovated and require fewer exposures to learn, and so would be more abundant in larger populations of speakers, where there are more people to generate innovations.

To test these hypotheses, agents in the authors' model encode a set of abstract "conventions", which they learn from other agents, but can also be probabilistically invented or forgotten. Conventions are divided into Easy conventions, analogous to lexical items, which can be learned with only one exposure, and Hard conventions, analogous to linguistic structures, which require two exposures to learn. Agents interact by producing conventions to convey to other agents, by sampling a convention from their inventory in proportion to how often each convention has been used or heard in the past, and by being exposed to conventions produced by other agents. Conventions in the system are considered "active" if they are known by a minimum number of agents in the population. The results of the model show that, in smaller populations of agents, the number of Easy and Hard active conventions is relatively equal, while in larger populations of agents, the number of Easy conventions is substantially greater than the number of Hard conventions. These results mirror the observed relationships, in natural language, between the number of speakers of a language and the degree of structural complexity, and vocabulary size.

As described here, modeling work using both iterated learning and interactive learning approaches has covered a range of topics in linguistics, and argued that at least some cross-linguistic typological tendencies may stem from pressures that arise through learning and communication between language users, rather than needing to be innately specified constraints on linguistic structures. These iterated and interactive learning models have largely been used independently, though some work (e.g. Dediu, 2009; Kirby et al., 2015; Smith et al., 2017) has argued that both types of dynamics are necessary to paint a complete picture of the influences that shape language learning and language change. One obstacle to implementing a model which integrates both types of agent interaction dynamics is that, to date, very little modeling work has been done which directly compares the effects of each model independently (see Kirby et al., 2015, for one example). In order to accurately identify and credit the source of a particular bias that might emerge in a combined model, it is necessary to understand the effects of each component model in isolation. Thus, one of the general goals of this dissertation is to provide a systematic exploration and comparison of the effects of the iterated learning model and interactive learning model, each applied independently to the same test cases, with the aim of establishing where the predictions of each model agree, and where they differ. As will be discussed in Chapters 2, 3, and 4, the biases which emerge from these models, when combined with MaxEnt grammars, agree in direction - away from gang effects and away from variation - differing primarily in the strength of these emergent biases.

In the remainder of this chapter, I describe the details of the specific implementations of the iterated learning model (§1.4.1) and the interactive learning model (§1.4.2) used throughout this dissertation, as well as how gradient typological predictions are derived from these models (§1.5).

1.4.1 Iterated learning model

The iterated learning model (see e.g. Kirby, 2001) is intended as a model of cultural transmission, in which a child agent learns from an adult model (a stable target grammar) for some period of time. Once the child agent's learning period is over, it becomes the adult model for a new child agent, and the process repeats across multiple generations of agents. In previous research, this learning dynamic has been combined with several different types of representational theories, such as connectionist networks (Hare & Elman, 1995), MaxEnt grammars (Staubs, 2014), and Bayesian models (see e.g. Rafferty et al., 2013).

In this dissertation, the iterated learning model is implemented as a computational agent-based learning model in which individual agents represent their linguistic knowledge using a MaxEnt grammar and modify their constraint weights in response to their training data using the update rule in (1.3). At each generation there are two agents, one serving as the teacher agent, and the other as the learner agent. Each agent in a simulation knows the constraint set, the set of possible inputs, and, for each input, the set of possible output candidates and their violation marks. Each new learner agent is initialized with some starting set of constraint weights, which depend on the parameter decisions for that simulation. At each step of the simulation, an input is sampled from the distribution over input forms. This work assumes a uniform probability distribution over input forms. Then, the teacher agent samples a corresponding output form from its MaxEnt grammar. The learner agent then samples an output form for the same input, using its own current MaxEnt grammar. The learner agent updates its MaxEnt grammar each time it produces an output form that differs from the output form it received from the teacher agent. At the end of the learning period, the learner agent replaces the teacher agent, and a new learner agent is introduced. The general structure of an iterated learning model simulation is given below in Algorithm 1.

The major parameters of this model that must be considered are the initializations of the initial teacher and the learner agents, the number of generations to iterate over, the number of learning steps allowed to each learner agent, and the learning rate for the update rule (1.3). As will be discussed in further detail for each test case,

Algorithm 1 General structure of an iterated learning simulation

```
initialize teacher
for each generation do
    initialize learner
    for each learning step do
        i = sample input form
        teacher form = teacher generate output form for i
        learner form = learner generate output form for i
        if teacher form != learner form then
            learner update grammar (see 1.3)
        end if
    end for
    teacher = learner
end for
```

two initialization conditions are used throughout this dissertation: one in which an agent's initial constraint weights yield equal probability on each competing output candidate, and one in which the initial constraint weights are randomly sampled. Because I have no principled reason to choose one initialization condition over the other, the results of both are presented in each case. The number of generations iterated over in each simulation is arbitrarily set to 50 generations, as in all cases this number is sufficient to observe trends in the data. The learning rate for the update rule (1.3) was likewise arbitrarily set to 0.1, though this choice is discussed further in §4.4.2. The number of learning steps, on the other hand, differed from case to case, as the size and complexity of each test case differed. In order to observe any biases that may emerge across generations, the number of learning steps allowed to each agent must be constrained. Otherwise, the learner agents conform to their teachers too closely, and there is no change between generations. As will be discussed further for each test case, the number of learning steps in each case was chosen by hand, by examining the trends in learning within one generation, from initial teacher to first generation of learners, and identifying a point before the learners begin to match the teachers too closely.

1.4.2 Interactive learning model

The interactive learning model is intended as a model of the influence of social interaction in language change. In the interactive learning model (see e.g. Dediu, 2009; Pater, 2012), there is no target grammar. Rather, two learner agents develop a shared grammar by exchanging data and learning from each other. In this work, a "shared grammar" refers to a state where both agents are highly likely to sample the same output form for each possible input form, not a state where both agents' grammars are identical.

In this dissertation, the interactive learning model is implemented as a computational agent-based learning model in which individual agents represent their linguistic knowledge using a MaxEnt grammar and modify their constraint weights in response to their training data using the update rule in (1.3). In each simulation there are two agents, who take turns playing the role of the teacher agent and the the learner agent. Each agent in a simulation knows the constraint set, the set of possible inputs, and, for each input, the set of possible output candidates and their violation marks. Each agent is initialized with some starting set of constraint weights, which depend on the parameter decisions for that simulation. At each step of the simulation, an input form is sampled from the set of possible input forms. This work assumes a uniform probability distribution over input forms. Then, each agent samples an output form for that input using their current MaxEnt grammars. If the current learner's form does not match the current teacher's form, then the learner agent updates its MaxEnt grammar. The general structure of an interactive learning model simulation is given below in Algorithm 2.

The major parameters of this model that must be considered are the initializations of the agents, the number of learning steps over which the agents interact, and the learning rate for the update rule (1.3). Just as discussed for the iterated learning model, two initialization conditions are used throughout this dissertation – one in

Algorithm 2 General structure of an interactive learning simulation

```
initialize Agent 1
initialize Agent 2
for each learning step do
    for each Agent do
        teacher = current Agent
        learner = other Agent
        i = sample input form
        teacher form = teacher generate output form for i
        learner form = learner generate output form for i
        if teacher form != learner form then
            learner update grammar (see 1.3)
        end if
    end for
end for
```

which an agent's initial constraint weights yield equal probability on each competing output candidate, and one in which the initial constraint weights are randomly sampled – and the learning rate for the update rule (1.3) was arbitrarily set to 0.1. The number of learning steps differed from case to case, as the size and complexity of each test case differed. The number of learning steps in each case was chosen by hand to be large enough to observe the emergence of a stable trend, without being unnecessarily large.

1.5 Generating a probabilistic typology

In order to explore the predictions of the iterated and interactive learning models when combined with MaxEnt grammars, the implementations of these models described above (§1.4.1 and §1.4.2) are applied to several test cases, as discussed in Chapters 2, 3, and 4. As defined in §1.3.3, the probabilistic typological predictions derived from these models are evaluated using two metrics: a distribution over possible pattern types, and a distribution over possible probability distributions, implemented as a measure of conditional entropy.

The probabilistic typological predictions of a given model on a given test case are generated by performing multiple runs of the model using the MaxEnt grammar space, yielding the predicted typological distribution. One "run" of the learning model consists of one instance of applying the methods detailed in Algorithm 1 for the iterated model and Algorithm 2 for the interactive model, which yields one possible learned MaxEnt grammar. Because the agents' productions and interactions are subject to random sampling, each run of the learning model could potentially have a different outcome. Thus, many runs of the model are performed, generating a distribution over possible learned grammars, which is then evaluated using the pattern type and conditional entropy metrics. The results of the models are compared to a baseline distribution, generated by simply sampling random grammars from the MaxEnt grammar space, without applying learning. Any differences between the predictions made by the model results and the baseline distribution can be attributed to the effect of learning.

The dissertation will proceed as follows. In Chapter 2 I discuss the results of applying these learning models to a simple, hypothetical typology, detailing how these models produce emergent learning biases towards deterministic grammars and away from gang effects. Chapter 3 investigates the bias away from cumulative constraint interactions in further detail, describing the results of applying the learning models to a more concrete example - a palatalization typology of contrast patterns between /s/ and /ʃ/ - while Chapter 4 provides more a more detailed examination of the bias towards more deterministic grammars, and investigates the distinction between variation and exceptionality. Chapter 5 provides discussion and conclusions.

CHAPTER 2
SIMPLE SYSTEM TEST CASE

In this chapter, I test the predictions of the iterated and interactive learning models using a simple MaxEnt grammar space as a minimal working example, which will allow for a detailed analysis of the models' inner workings and the results of the interaction between grammar and learning. The results of both models show a bias away from gang effects, in which multiple violations of lower-weighted constraints can cumulatively outweigh violations of higher-weighted constraints, and a bias away from variation in the grammar, towards more deterministic patterns in which probability is accumulated on one output candidate over its competitors. The bias away from gang effects is explored further in Chapter 3, and the bias away from variation is explored further in Chapter 4.

2.1 Introduction

In order to systematically explore how the iterated and interactive learning models work, and what kinds of effects they have on the predicted gradient typologies, I constructed a small grammar to serve as a minimal working example, so that the results of the models might be as transparently interpretable as possible. In this simple, hypothetical system, there are two constraints, ConX and ConY, and two input forms, I_1 and I_2. For each input, there are two competing output candidates: [A] and [B] are competing outputs for input I_1, and [C] and [D] are competing outputs for I_2. This system is illustrated in Tableau 2.1.

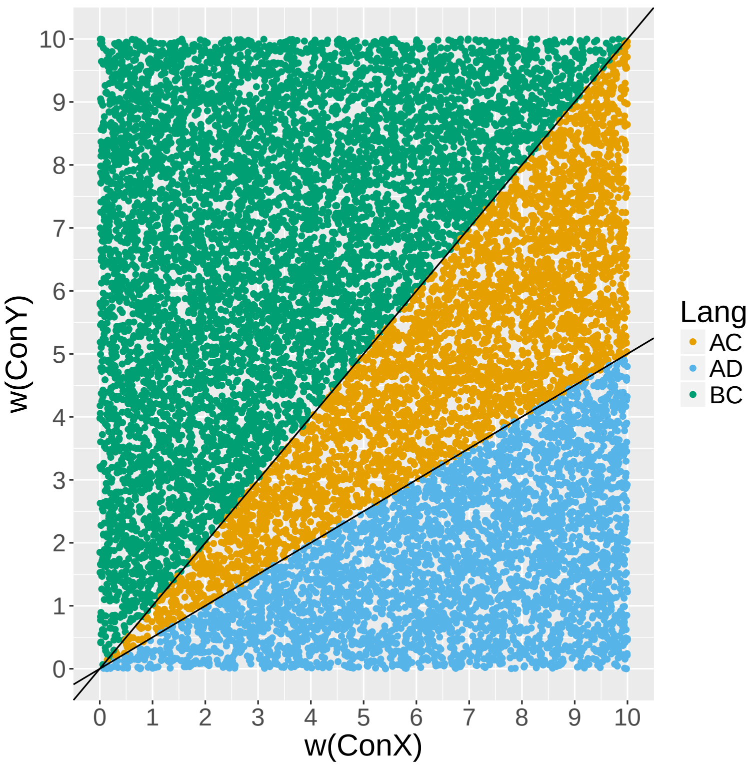
Tableau 2.1. Tableaux showing the simple test grammar

$/I_1/$	ConX	ConY
A		-1
B	-1	

$/I_2/$	ConX	ConY
C	-1	
D		-2

The advantage of using this simple two-constraint system as a first test of the models' effects is that it allows for straightforward analysis, and straightforward visual representations, of the evolution of the agents' grammars as learning progresses. This simple system can be represented by a two-dimensional plane, as graphed in Figure 2.1, where the x-axis plots the weight of constraint ConX and the y-axis plots the weight of constraint ConY.[1]

Figure 2.1. Plot of randomly sampled weight pairs in the simple test grammar space, color coded by pattern type



[1] As constraint weights are restricted to positive values, so the axes are restricted to show only positive values.

All possible grammars in this system can then be plotted on this plane using the constraint weights as a coordinate pair ($w(\textsc{ConX})$, $w(\textsc{ConY})$). The plane can additionally be divided into regions differentiating weight pairs which give higher probability to one output candidate over its competitor(s). The borders between these regions for the simple system are plotted in Figure 2.1 as lines along which lie all weight pairs which yield equal probability to the two pairs of competing output candidates, [A] vs. [B] and [C] vs. [D]. The line separating weight pairs which give higher probability to candidate [A] from weight pairs which give higher probability to candidate [B] is given by the equation $H(A) = H(B)$, or $w(\textsc{ConX}) = w(\textsc{ConY})$. Above this line, where $w(\textsc{ConY}) > w(\textsc{ConX})$, candidate [B] has higher probability and below this line, where $w(\textsc{ConX}) > w(\textsc{ConY})$, candidate [A] has higher probability. The line separating weight pairs which give higher probability to candidate [C] from weight pairs which give higher probability to candidate [D] is given by the equation $H(C) = H(D)$, or $w(\textsc{ConX}) = 2w(\textsc{ConY})$. Above this line, where $2w(\textsc{ConY}) > w(\textsc{ConX})$, candidate [C] has higher probability and below this line, where $w(\textsc{ConX}) > 2w(\textsc{ConY})$, candidate [D] has higher probability.

As can be seen in Figure 2.1, these lines of equal probability divide the plane into three regions, corresponding to the three possible patterns in this simple system. I use "pattern" to refer to a unique set of output candidates such that each has the highest probability among its set of competitors. These three possible patterns, labeled by the set of highest-probability output candidates, are listed in Table 2.1 with the weighting conditions which yield them.

Table 2.1. The set of possible patterns in the simple system typology

	Pattern	Weighting condition
i.	BC	$w(\textsc{ConY}) > w(\textsc{ConX})$
ii.	AD	$w(\textsc{ConX}) > 2w(\textsc{ConY})$
iii.	AC	$2w(\textsc{ConY}) > w(\textsc{ConX}) > w(\textsc{ConY})$

The weighting condition which yields the BC pattern, illustrated in Tableau 2.2[2], is w(ConY) > w(ConX); that is, when the weight of constraint ConY is greater than the weight of constraint ConX. Under this weighting condition, the output candidates which violate constraint ConY ([A] and [D]) receive greater penalty than the candidates which violate constraint ConX ([B] and [C]), making the probabilities assigned to [B] and [C] higher than those given to [A] and [D].

Tableau 2.2. Tableaux showing an example of the BC pattern in the simple system

	1	4	H	p
/I_1/	ConX	ConY		
A		-1	-4	0.047
☞B	-1		-1	0.953

	1	4	H	p
/I_2/	ConX	ConY		
☞C	-1		-1	0.999
D		-2	-8	0.001

The weighting condition which yields the AD pattern, illustrated in Tableau 2.3, is w(ConX) > 2w(ConY); that is, when the weight of constraint ConX is greater than two times the weight of constraint ConY. It is necessary that the weight of ConX be greater than two times ConY, and not simply greater than ConY, because output [D] violates constraint ConY twice ($H(D) = 2w$(ConY)), where [C] violates ConX only once ($H(C) = w$(ConX)). In order for [D] to be more harmonic than [C], then, the penalty given to a single violation of constraint ConX must be greater than the penalty assigned to two violations of constraint ConY. Under this weighting condition, the output candidates which violate constraint ConX ([B] and [C]) receive greater penalty than the candidates which violate constraint ConY ([A] and [D]), making the probabilities assigned to [A] and [D] higher than those given to [B] and [C].

The third possible pattern in this system, the AC pattern, is the result of a gang effect between the constraints, as illustrated in Tableau 2.4. The weighting condition

[2] This and all following patterns are illustrated with a set of weights which yield a fairly deterministic grammar, such that each candidate is assigned >95% or <5% probability.

Tableau 2.3. Tableaux showing an example of the AD pattern in the simple system

	5	1	H	p
$/I_1/$	ConX	ConY		
☞A		-1	-1	0.982
B	-1		-5	0.018

	5	1	H	p
$/I_2/$	ConX	ConY		
C	-1		-5	0.047
☞D		-2	-2	0.953

which yields the AC pattern is $2w(\textsc{ConY}) > w(\textsc{ConX}) > w(\textsc{ConY})$, that is, when twice the weight of constraint ConY is greater than the weight of constraint ConX, and the weight of constraint ConX is greater than the weight of constraint ConY. This is because, in order for candidate [A] to be better than [B], the penalty given to a single violation of ConX must be greater than the penalty assigned to a single violation of ConY, but in order for candidate [C] to be better than [D], the penalty assigned to two violations of ConY must be greater than the penalty given to a single violation of ConX.

Tableau 2.4. Tableaux showing an example of the AC pattern in the simple system

	9	6	H	p
$/I_1/$	ConX	ConY		
☞A		-1	-6	0.953
B	-1		-9	0.047

	9	6	H	p
$/I_2/$	ConX	ConY		
☞C	-1		-9	0.953
D		-2	-12	0.047

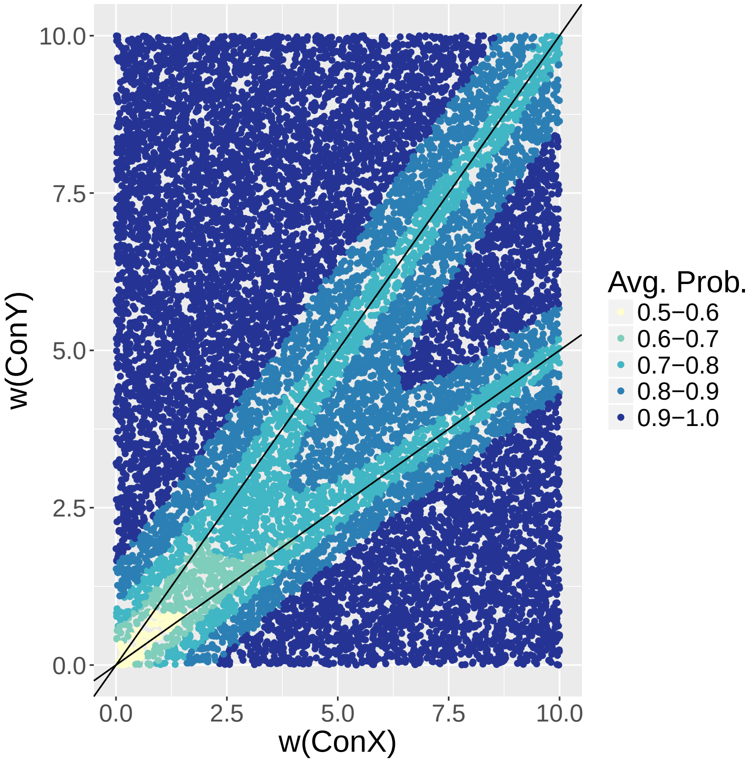
Before evaluating the effects of the iterated and interactive learning models on this system, it is important to establish the properties of the system by itself, and any pre-existing biases that may influence the distribution over patterns or the rate of variability predicted typologically. This is necessary in order to be able to evaluate whether the distributions predicted by the learning models are the result of biases introduced by the learning mechanics, or whether the learning models are simply reproducing biases already present in the grammatical system. The distribution over patterns which emerges without applying learning can be established by estimating the proportion of weight pairs $(w(\textsc{ConX}), w(\textsc{ConY}))$ which yield each of the three possible pattern types. This is similar in principle to counting the proportion of total constraint rankings which yields a particular pattern type in a ranked-constraint grammatical model (r-volume, see Riggle, 2010). A probabilistic typological prediction calculated in this way assumes that the frequency of a pattern is directly proportional to the number of weight pairs or rankings which produce that pattern, with no intervening influence.

Estimating biases concerning variability and possible probability distributions is somewhat more difficult, because there are in principle an infinite number of possible probability distributions. In order to simplify the task, I binned the range of possible probability distributions into five categories: $p = 0.5 - 0.6$, $p = 0.6 - 0.7$, $p = 0.7 - 0.8$, $p = 0.8 - 0.9$, and $p = 0.9 - 1.0$. I assigned each probability distribution to a category by calculating the average probability over the highest probability candidates in each set of competitors. For example, in Tableau 2.3, the highest probability candidates from each set of competitors are [A] ($p = 0.982$) and [D] ($p = 0.953$). Their average probability is $p = 0.968$, so this grammar would be assigned to the 0.9-1.0 category.

In order to estimate the predictions for the distribution over patterns and probabilities without learning, 15,000 weight pairs were sampled from a uniform distribution with the range 0.0-10.0.[3] These points are plotted in Figure 2.1, where they are color coded by pattern type, and in Figure 2.2, where they are color coded by probability. The distribution over patterns and probabilities is given in Table 2.2. The BC pattern occupies half of the weight space, while the AD and AC patterns each occupy one quarter of the weight space. A majority of sampled weight pairs (61%) yield a highly categorical grammar (average probability between 0.9-1.0). The probability of sampling a more variable grammar decreases as average probability decreases. As can be seen in Figure 2.2, the weights that produce more variable grammars lie closer to

[3] The upper bound on the sampling distribution was arbitrarily chosen. There is no theoretical upper bound on the value of constraint weights. Sampling a larger range produces the same results for the distribution over pattern types, and would increase the skew towards more categorical patterns.

Figure 2.2. Plot of randomly sampled weight pairs, color coded by average highest probability divided into five bins



the borders between pattern types, where the pairs of competing output candidates have equal probability. The most variable grammars are clustered near the origin, where the borders between the pattern types approach each other.

Table 2.2. The sampled baseline distribution over pattern types and average highest probabilities, divided into five bins, for the simple system typology

	Pattern			Average Highest Probability				
	BC	AC	AD	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Proportion	0.50	0.25	0.25	0.01	0.03	0.14	0.22	0.61

Figure 2.2 also hints at an interaction between pattern type and average probability. The region corresponding to the gang effect AC pattern, which lies between the BC pattern (top left) and the AD pattern (bottom right), visually contains most of the sampled weight pairs which yield more variable grammars. This visual impression is confirmed in Table 2.3, which shows the proportion of the sampled grammars

belonging to each probability bin that correspond to each pattern type. For example, among the sampled weight pairs which yield an average highest probability between 0.5-0.6, 57% of those correspond to the gang effect AC pattern. Note that these are proportions out of the proportion of grammars corresponding to each probability bin, so that is 57% of the 1% of grammars falling into the 0.5-0.6 range in the baseline distribution.

Table 2.3. The proportion of sampled weight pairs in the baseline distribution corresponding to each pattern type, normalized by bin of average highest probability

		Prop. of	Pattern Type		
		Baseline	BC	AC	AD
Avg. Prob.	0.5-0.6	0.01	0.12	0.57	0.32
	0.6-0.7	0.03	0.21	0.52	0.27
	0.7-0.8	0.14	0.31	0.50	0.18
	0.8-0.9	0.22	0.37	0.44	0.19
	0.9-1.0	0.61	0.60	0.11	0.29

In all but the highest probability bin, the majority of weight pairs yield the AC pattern. As average probability increases, this skew towards the AC pattern decreases; the proportion of weight pairs corresponding to the AD pattern decreases as well, while the proportion corresponding to the BC pattern increases. The difference between the highest probability bin (0.9-1.0) and the others is rather marked; among the most deterministic grammars, weight pairs yielding the BC pattern are the majority by a rather large margin, while the AC pattern becomes the minority.

In sum, when the space of possible grammars is conceptualized as a multidimensional space of positive weight values, where each dimension corresponds to one constraint, it is possible to investigate the properties of the typological space. In this simple typology, the region corresponding to the BC pattern is the largest, and contains both the smallest proportion of variable grammars and the largest proportion of deterministic grammars. While the regions corresponding to the AD pattern and the gang effect AC pattern are equal in size, the AC pattern's region contains more variable grammars and fewer deterministic grammars than the AD pattern's region.

The remainder of this chapter describes the results of the learning model simulations performed with the simple typology. §2.2 reports and discusses the effects of simulations with the interactive learning model, while §2.3 reports and discusses the effects of simulations with the iterated learning model. As I will show, both models display (1) a robust bias away from the gang effect AC pattern, where agents learn the AC pattern less often than would be expected by chance, and (2) a tendency towards more deterministic grammars, where agents tend to accumulate probability on one output candidate over its competitor, with the primary difference between the models being the underlying mechanisms that drive their behavior. §2.4 provides an overall discussion of the results of the interactive and iterated learning models, and their implications for empirical phonological typology.

2.2 Interactive learning model simulations

In this section, I present the results of applying the interactive learning model (see §1.4.2) to the simple typology. In the interactive learning model, two learner agents exchange data between each other, developing a shared grammar through mutual imitation, with no fixed target grammar. I test two initialization conditions: one in which both agents are initialized with constraint weights at zero (the zero-weight initialization condition), and one in which both agents are initialized with random constraint weights (the random-weight initialization condition; subject to a balancing constraint as described below).

In the zero-weight initialization condition, each learning agent was initialized with all constraint weights at zero. This gives the learning agents an unbiased starting point; with all weights at zero, the set of competing output candidates for each input have equal probability ($p = 0.5$). 1,500 runs with this initialization were performed.

In the random-weight initialization condition, both of the learning agents were initialized with the same set of constraint weights, which were randomly sampled from a uniform distribution with range 0.0-10.0, subject to a balancing constraint which ensured that an equal number of runs were initialized in each pattern, and in each probability bin. This balancing of the sampled initial states was done in order to avoid any potential confounds from the pre-existing skews in the grammar space. 7,500 runs with this initialization were performed, with 500 runs initialized in each combination of pattern type and probability category.

For simulations in both initialization conditions, the learning agents exchanged data for 5,000 learning steps, with a learning rate of 0.1. As I will show, the results of these simulations, as well as the iterated learning simulations in §2.3, demonstrate a bias away from the gang effect AC pattern, as well as a bias away from variability in the grammar.

2.2.1 Interactive learning simulation results

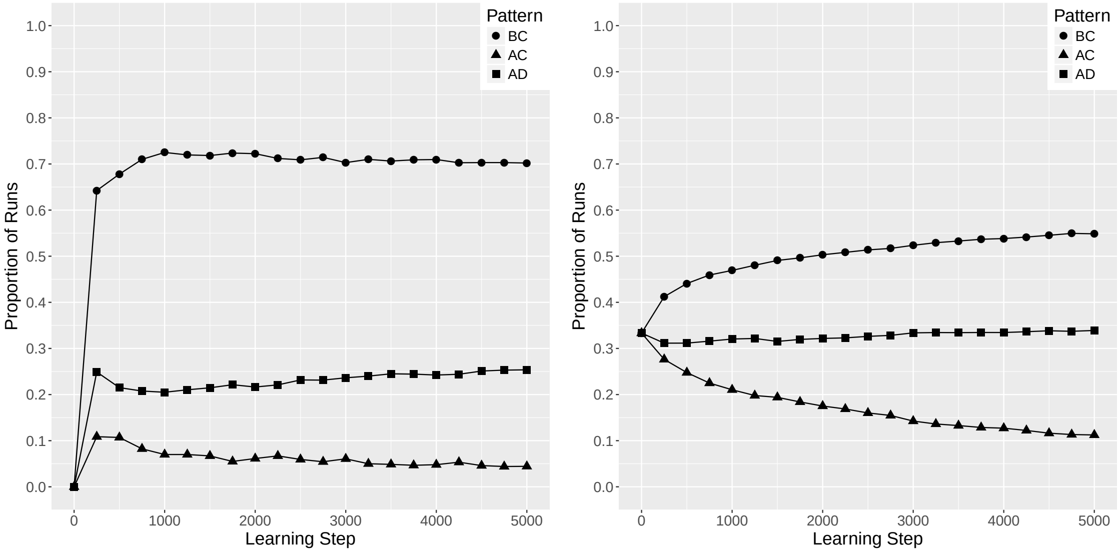
Table 2.4 summarizes the agents' distribution over pattern types and probability categories at the end of the simulations. The table shows results for both initialization conditions, as well as repeating the sampled baseline distribution. The "Equal Distribution" shows the starting distribution of the random-weight initialization condition, in which the sampled initial states are balanced across pattern types and probability bins. At this point in learning, the agents in both initialization conditions demonstrate several biases. The first is a bias towards more categorical patterns, with agents in a large majority of runs assigning an average probability between 0.9-1.0 to the highest probability output candidates. This bias is more extreme than present in the sampled baseline distribution, with 81% of runs in the zero-weight condition and 84% of runs in the random-weight condition falling into this probability bin, compared to 61% in the sampled baseline and a mere 20% in the equal distribution. Another is a bias away from the gang effect AC pattern, with agents assigning highest probability to candidates [A] and [C] in fewer runs than occur by chance in the sampled baseline (25%, versus 4% in the zero-weight condition and 11% in the random-weight condition), and an even bigger difference between the equal distribution and the results of the random-weight condition (33% vs 11%). The differences in the results between initialization conditions is linked to the shape of the baseline distribution (see Figures 2.1 and 2.2), and the directions in which the agents tend to drift, as discussed in §2.2.2.

Table 2.4. The distribution over pattern types and average highest probabilities after learning with the interactive learning model in the simple test typology

	Pattern			Average Highest Probability				
	BC	AC	AD	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Equal Distribution	0.33	0.33	0.33	0.20	0.20	0.20	0.20	0.20
Sampled Baseline	0.50	0.25	0.25	0.01	0.03	0.14	0.22	0.61
Zero-weight Cond.	0.70	0.04	0.25	0.01	0.03	0.07	0.09	0.81
Random-weight Cond.	0.55	0.11	0.34	0.01	0.01	0.06	0.10	0.84

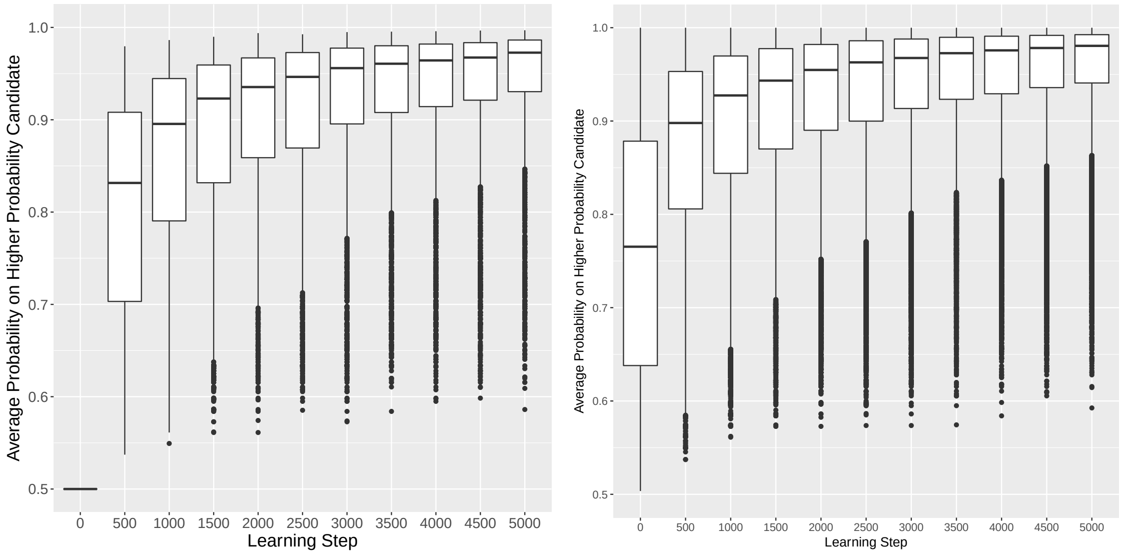
Taking a closer look at the distribution over pattern types, the graphs in Figure 2.3 plot the proportion of runs in which the learning agents hold each pattern across learning steps, tracking the change in the distribution over patterns as the simulation progresses. The left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. For the zero-weight condition, at zero learning steps, the agents are still in the initial state with equal probability on all output candidates, and thus are not categorized as belonging to any pattern. Once the learning agents begin exchanging data, the distribution over pattern types quickly moves into a relatively stable state, with agents in most runs learning the BC pattern, and the least learning the AC pattern. As the simulation approaches 5,000 learning steps, the distribution over patterns shows only very small changes. For the random-weight initialization condition, at zero learning steps, the agents are still in the initial state, with an equal proportion of runs corresponding to each pattern type and probability bin. As the learning agents exchange data, the distribution over pattern types gradually shifts towards a distribution similar to, but less extreme than that seen in the zero-weight condition. More learning agent pairs are moving into the BC pattern, and more are moving out of the AC pattern, while the number of agents in the AD pattern remains largely stable.

Figure 2.3. Plots showing the change in the distribution over pattern types across simulation time, using the interactive learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)



Taking a closer look at the distribution over probabilities, the graphs in Figure 2.4 plot the average probabilities assigned to the highest probability output candidates across learning steps, tracking the change in the distribution over probabilities as the simulation progresses. Again, the left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. As the learning agents exchange data in both conditions, they gradually accumulate a greater majority of probability onto the highest probability candidates, as seen in both the increase in median probability and the decrease in variance across the distributions.

Figure 2.4. Plot showing the change in the distribution over the average highest probability across simulation time, using the interactive learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)



2.2.2 Interactive learning discussion

The above sections describe the results of applying the interactive learning model to the simple test system under two initialization conditions: one in which both agents are initialized with constraint weights at zero, and one in which both agents are initialized with randomly sampled constraint weights, where the samples were balanced across runs for both pattern type and probability distribution. As shown in the results for both initialization conditions, the interactive learning model demonstrates a bias towards more categorical patterns, and a bias away from the AC pattern.

The key to understanding these biases, and their robustness across initialization conditions, lies in the core mechanics of agent interaction and error-driven learning. In the interactive learning model, the agents take turns playing the roles of teacher and learner, and the agent in the learner role only updates its grammar when its sampled output form does not match the sampled output form of the agent playing the teacher role. When the agents have more variable grammars, they are less likely to sample the same output form, and thus are more likely to make an update to their grammars. When the agents have more deterministic grammars, they are more likely to sample the same output form (provided the agents' grammars correspond to the same pattern), and thus are less likely to make an update to their grammars. Pairs of agents with more variable grammars, then, show a faster rate of change than agents with more deterministic grammars, who are relatively more likely to preserve the same grammar across multiple learning steps.

The interactive learning model's bias away from variable grammars is linked to the model's bias away from the gang effect AC pattern. As seen in Table 2.2, the proportion of possible grammars corresponding to more variable patterns is smaller than the proportion of possible grammars corresponding to more deterministic patterns. As the constraint weights get farther from zero, and farther from the borders between pattern types, the probability distribution generated with those weights becomes more peaked, with probabilities approaching $p = 0$ or $p = 1$. Looking once again at Figure 2.1, it can be seen that region of constraint weight pairs which generate the gang effect AC pattern lies in the center of the plane, bordered on either side by the other patterns, BC and AD. Because the AC pattern borders two regions, compared to one each for BC and AD, proportionally more of the constraint weight pairs in the region corresponding to the AC pattern also correspond to more variable grammars. The model's bias away from variable grammars, then, reduces the predicted probability of the gang effect AC pattern relative to the BC and AD patterns.

The difference between the results for each initialization condition regarding the strength of the bias away from the AC pattern can additionally be explained by considering the bias away from variation. The predicted probability of the AC pattern was lower in the zero-weight initialization condition than in the random-weight initialization condition. Constraint weights closer to zero, and closer to the borders between pattern types, yield more variable grammars; additionally, because the borders between the pattern types approach each other as they approach the origin, the lowest constraint weight pairs which yield a deterministic gang effect AC pattern lie farther from zero than the lowest weight pairs which yield a deterministic BC or AD pattern. Agents in the zero-weight initialization condition, then, are more likely to end up in a deterministic BC or AD pattern because those regions are closer to their starting point.

2.3 Iterated learning model simulations

In this section, I present the results of applying the iterated learning model (Kirby, 2001, see §1.4.1) to the simple typology. In each generation of the iterated learning model, the child (learner) agent learns from an adult (teacher) model (a stable target grammar) for some period of time. Once the child agent's learning period is over, it becomes the adult model for a new child agent, and the process repeats across multiple generations of agents. In the interactive learning model results in the previous section, each run of the simulation consisted of two peer learner agents, who performed the same actions and were subjected to the same initialization constraints; however, in the iterated learning model, each agent goes through a learner stage, as a child, and a teacher stage, as an adult, and each run of the simulation consists of a multi-agent chain with some initial adult target grammar provided for the first agent. The application of the zero-weight and random-weight initialization conditions were thus adapted to better fit this difference in the structure of the model. New child agents introduced into the chain were subject to one of two initialization conditions: one in which the new learner agents are initialized with constraint weights at zero, and one in which new learner agents are initialized with random constraint weights, sampled from a uniform distribution with the range 0.0-10.0, with no balancing constraint. The initial target grammars in each chain were randomly sampled, subject to a balancing constraint across runs so that an equal number of chains (500) were initialized in each combination of pattern type and probability bin, for a total of 7,500 runs in each condition. The learning rate was 0.1.

The results here are presented in two sections, one showing the learning paths of a single generation of agents learning from their fixed target grammars for 5,000 learning steps, and the second showing the results of iterating over multiple generations of agents, who each learn from their target grammar for 1,000 learning steps. As I will show, the results of these simulations, as was the case with the interactive learning simulations in §2.2, demonstrate a bias away from the gang effect AC pattern, as well as a bias away from variability in the grammar.

2.3.1 Simulation results: within one generation

This section presents the results of transmitting the balanced set of randomly sampled initial states from the initial teacher to the first generation of learners. This is useful because understanding the trends observed in the transmission of a pattern from teacher to learner will aid in understanding the trends observed in the transmission of patterns across multiple generations of agents. In the results presented below, the learner agents received data from their target grammars for 5,000 learning steps.

Table 2.5 shows the final distributions over pattern types and probability categories across runs for the learner agents in both initialization conditions, as well as giving the sampled baseline distribution and the target distribution for comparison. Due to the balanced sampling for the grammars of the teacher agents, the target distribution is divided evenly across pattern types and across probabilities. As can be seen in comparing the results for the zero-weight and random-weight initialization conditions, the distribution over patterns and probabilities after 5,000 learning steps is the same, suggesting that 5,000 learning steps in this system is enough to obscure any influence of the initial state. In these simulations, the gang effect AC pattern is learned more often than expected from the target distribution, and the BC and AD patterns are learned less often than expected, though the proportions are not drastically different from the target distribution. This is contrary to the results found in the interactive learning model, which displayed a bias away from the gang effect AC pattern. On the other hand, a bias towards more categorical patterns does emerge in these results, though it is weaker than in the interactive learning model. The learner agents in Table 2.5 learned grammars in the two lowest probability categories less often than expected from the target distribution, and grammars in higher categories more often than expected.

Table 2.5. The distribution over pattern types and average highest probabilities after one generation of learning with the iterated learning model in the simple test typology

	Pattern			Average Highest Probability				
	BC	AC	AD	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.50	0.25	0.25	0.01	0.03	0.14	0.22	0.61
Target Distribution	0.33	0.33	0.33	0.20	0.20	0.20	0.20	0.20
Zero-weight Cond.	0.31	0.38	0.31	0.10	0.19	0.25	0.23	0.23
Random-weight Cond.	0.31	0.38	0.31	0.10	0.19	0.25	0.23	0.23

The graphs in Figure 2.5 take a closer look at the distribution over pattern types, plotting the proportion of runs in which the learner agents hold each pattern type across learning steps. The dashed line represents the target distribution, in which each pattern type is equally represented (1/3 ≈ 0.33). The left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. At zero learning steps, the learner agents are in their initial state; in the zero-weight initialization condition, this means weights at zero and thus no preference for any output forms, and in the random-weight initialization condition, this produces the distribution seen in the sampled baseline. As learning begins, the learner agents in the zero-weight initialization condition initially favor the BC and AD patterns, with a higher proportion of runs moving into the gang effect AC pattern starting around 1,500 learning steps. In the random-weight initialization condition, on the other hand, there is a large initial increase in the proportion of runs corresponding to the AC pattern. This initial increase can be attributed to cases in which the learner agent is initialized in a pattern different than that of its target grammar; in these cases, the learner agent will always pass through a stage where it holds a set of weights that corresponds to the AC pattern, because the AC pattern's region lies between the BC and AD regions. The proportion of runs in the AC pattern gradually decreases until the distribution shows the same slight overrepresentation of the AC pattern seen in the zero-weight initialization condition. In both conditions, the distribution over patterns appears to remain fairly stable after around 2,000 learning steps.

Figure 2.5. Plots showing the change in the distribution over pattern types across one generation of learning with the iterated learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)

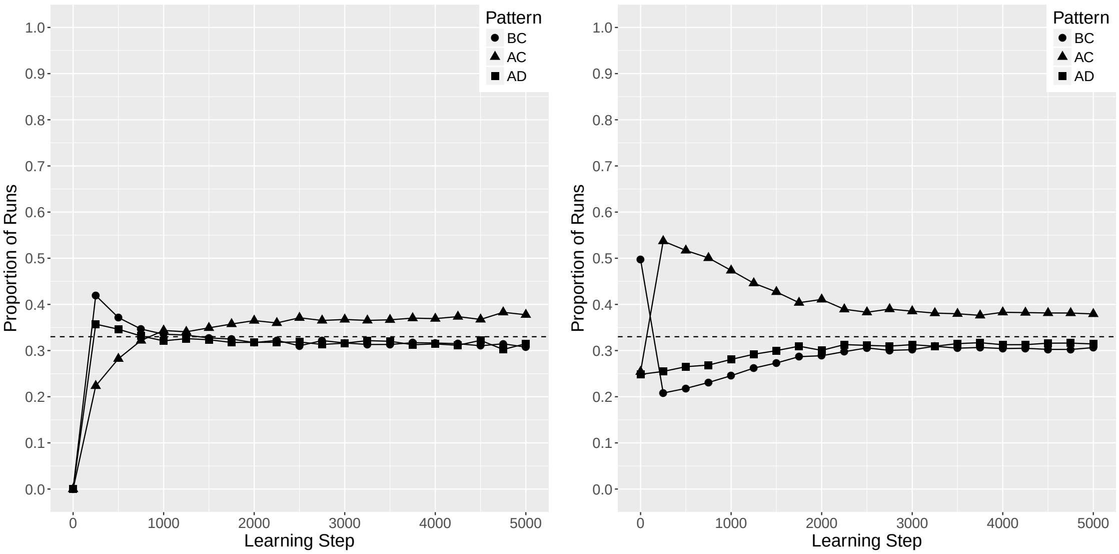
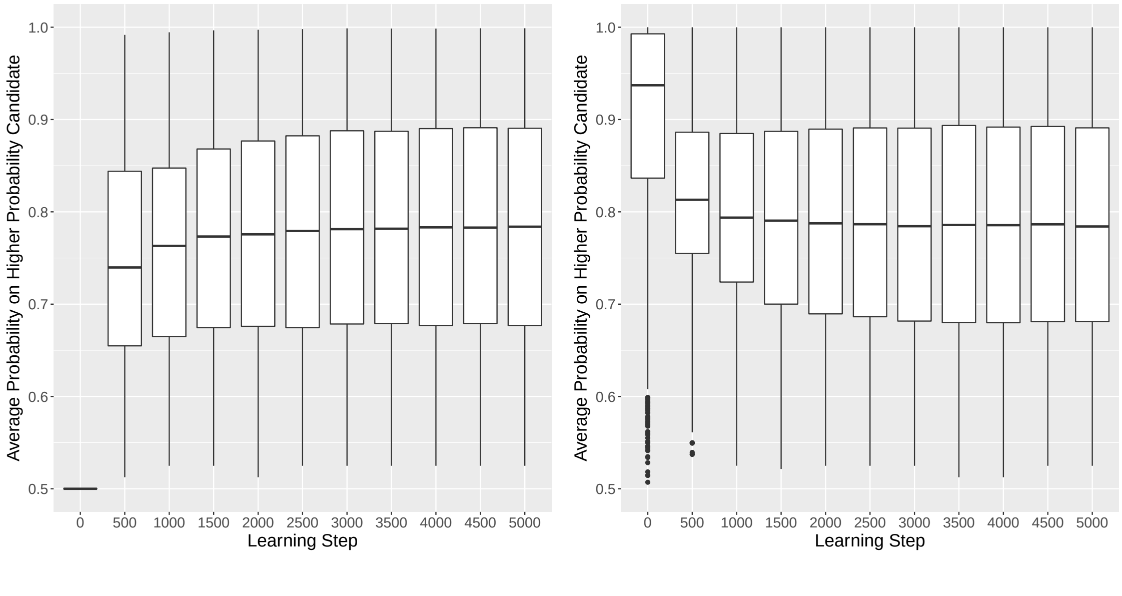


Figure 2.6 takes a closer look at how the average probability the learner agents assign to the highest probability output candidates changes throughout learning, plotting the distribution over average probabilities by learning step. Again, the left-hand

graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. As learning progresses, the distribution over probabilities increases to match the distribution over the target grammars, which is balanced across the probability categories. In the zero-weight initialization condition, the average probabilities assigned by the learning agents increases to match the target distribution; this is because, at zero learning steps, the agents assign equal probability among competing output candidates, and then gradually increase their constraint weights to fit their target grammar. In the random-weight initialization condition, on the other hand, the distribution over average probabilities assigned by the learning agents decreases to match the target distribution. This is because, as seen in the baseline distribution, the randomly sampled initial states for the learning agents contain a greater proportion of deterministic grammars than present in the balanced target distribution.

Figure 2.6. Plots showing the change in the distribution over average highest probabilities across one generation of learning with the iterated learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)



2.3.2 Simulation results: across generations

This section shows the results of iterating learning over 50 generations, where each new learner agent learns from its target grammar for 1,000 learning steps. As seen above in the presentation of the simulation results within one generation, 5,000 learning steps is enough for the learning agents in this system to reach a relatively stable distribution over pattern types and probabilities. Iterating over multiple generations of agents, when each generation is able to reach this stable point, results in little to no change across generations. In order to have a chance to observe any biases at all emerge from this model, the agents must be cut off at a point earlier in learning. Because the distributions over patterns, seen in Figure 2.5, appear to become relatively stable around 2,000 learning steps, the learning agents in the simulations presented in this section were given 1,000 learning steps: the midway point between the initial state and the stable state.

Table 2.5 shows the distributions over pattern types and probabilities after 50 generations for both initialization conditions, as well as giving the sampled baseline distribution and the initial target distribution for comparison. In both initialization conditions, the smallest proportion of runs at this point in the simulation correspond the gang effect AC pattern, followed by the AD pattern, then the BC pattern, which also represented the largest proportion of the baseline distribution. The differences between the proportions are more extreme in the zero-weight initialization condition than in the random-weight initialization condition; this difference between the initialization conditions is linked to differences in potential directions for mislearning, as discussed in §2.3.3. Both conditions additionally show a shift towards more deterministic patterns, with nearly three-quarters of the runs belonging to the highest probability bin.

The graphs in Figure 2.7 take a closer look at the distribution over pattern types, plotting the proportion of runs in which the learner agents hold each pattern type

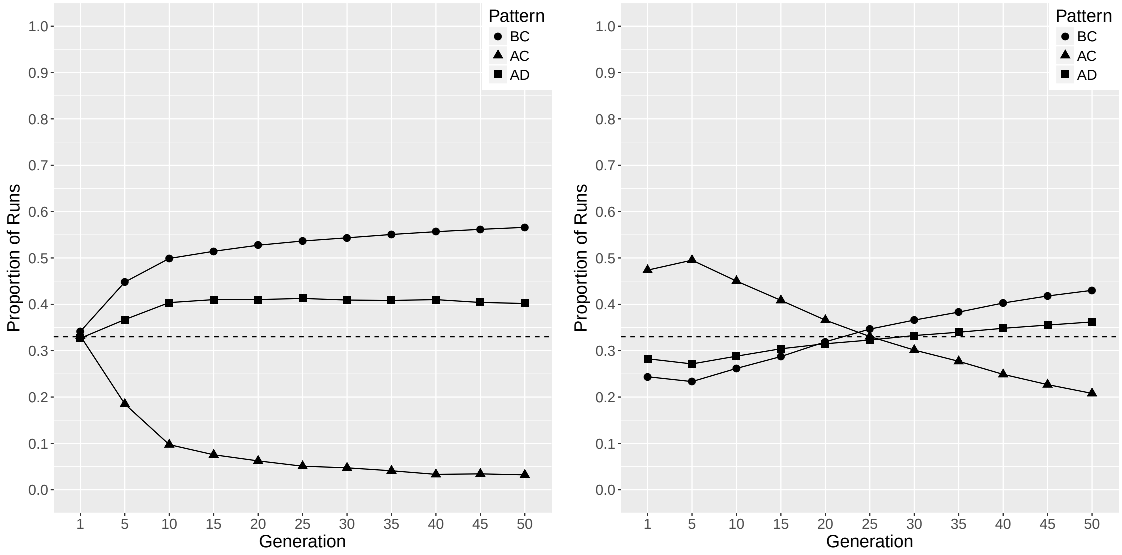
Table 2.6. The distribution over pattern types and average highest probabilities after 50 generations of learning with the iterated learning model in the simple test typology

	Pattern			Average Highest Probability				
	BC	AC	AD	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.50	0.25	0.25	0.01	0.03	0.14	0.22	0.61
Target Distribution	0.33	0.33	0.33	0.20	0.20	0.20	0.20	0.20
Zero-weight Cond.	0.57	0.03	0.40	0.01	0.05	0.06	0.14	0.74
Random-weight Cond.	0.43	0.21	0.36	0.00	0.01	0.10	0.20	0.71

across learning steps. Again, the dashed line represents the target distribution, in which each pattern type is equally represented ($1/3 \approx 0.33$), the left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. In the zero-weight condition, the proportion of runs corresponding to the BC or AD patterns rises quickly across the first ten generations, while the proportion of runs corresponding to the gang effect AC pattern decreases quickly. The rate of change across generations then decreases, with the proportion of BC increasing gradually while the proportions of AD and AC decrease. In the random-weight condition, the proportion of runs corresponding to the gang effect AC pattern is initially highest; this proportion then decreases as the simulation iterates over subsequent generations, eventually crossing below the other patterns. At the final generation in the simulation, the AC pattern is present in the fewest runs, while the BC pattern is present in the most. The slope of the lines plotting the proportions indicates that they would likely continue to grow farther apart, if the simulation were allowed to continue running.

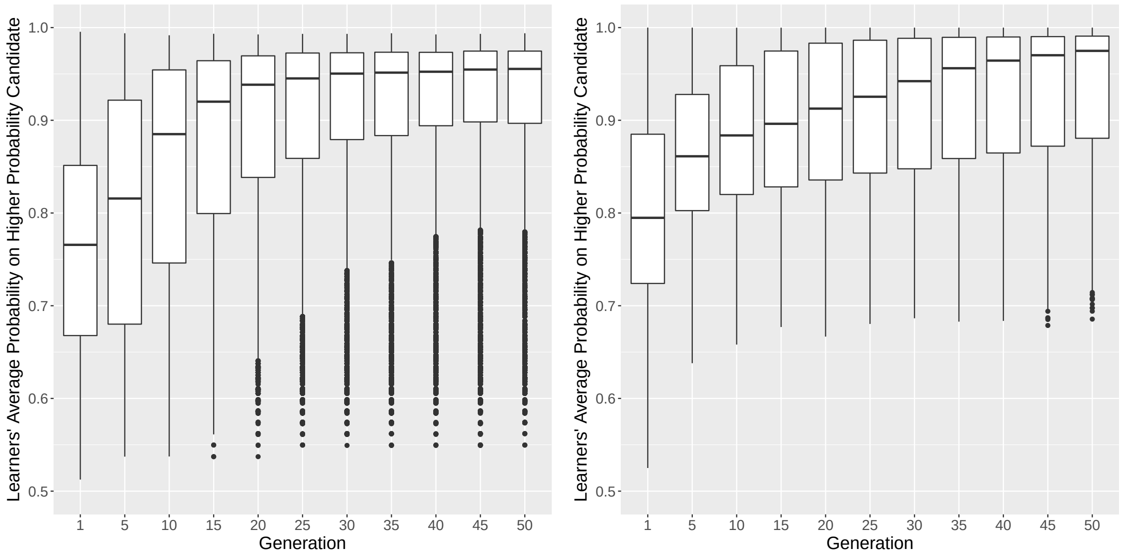
The graphs in Figure 2.8 plot the change in the distribution over the probabilities assigned to higher probability candidates across generations of learner agents (zero-weight condition results are on the left, and random-weight condition results are on the right). As generation increases, the mean average probability increases, and the variance of the distribution over probabilities decreases, showing that the bias

Figure 2.7. Plots showing the change in the distribution over pattern types across 50 generations of learning with the iterated learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)



towards more deterministic grammars is compounded as the grammars are iterated over multiple generations of agents.

Figure 2.8. Plots showing the change in the distribution over average highest probabilities across 50 generations of learning with the iterated learning model in the simple test typology (Left: zero-weight condition, Right: random-weight condition)



2.3.3 Iterated learning discussion

The above sections describe the results of applying the iterated learning model to the simple test system, looking at both the effects within one generation (i.e. from initial teacher to first learner) and the effects of iterating across a number of generations. Two initialization conditions were considered: one in which new learner agents were initialized with constraint weights of zero, and one in which new learner agents were initialized with constraint weights randomly sampled from a uniform distribution with the range 0.0-10.0. In all cases, the first target grammars in each chain were generated by random sampling, and were balanced across runs so that an equal number of chains (500) were initialized in each combination of pattern type and average probability bin, making 7,500 total runs in each condition.

Within one generation, from one teacher agent to the next learner, the learning agents are tasked with gradually increasing their fit to their target pattern as learning progresses. In the simulations presented here, the distribution across runs slightly overrepresented the gang effect AC pattern as compared to the target distribution, which contained an equal proportion of each pattern type (see Figure 2.5). This overrepresentation occurs because the gang effect AC pattern acts as an attractor state for all but the most deterministic grammars.

As was observed in the sampled baseline distribution (see Table 2.3), the majority of possible grammars in all but the highest probability bin also belong to the region corresponding to the AC pattern. Because the sampled target grammars were balanced across pattern type and average probability, the learning agents in most runs of the simulation were more likely to mislearn their target pattern as the AC pattern, by chance. This trend is confirmed by examining the transition probabilities; the left-hand table in Table 2.7 shows the proportion of runs in which the learning agent learned each pattern type, given a target grammar corresponding to a particular combination of pattern type and average probability. In all cases[4], the BC and AD patterns are being mislearned as the AC pattern more often than the AC pattern is being mislearned as the BC or AD patterns. The right-hand table in Table 2.7 condenses this information to be more easily visible; for the grammars in each target average probability range, it shows the proportion of learners who incorrectly learned the AC pattern, given that their target pattern was BC or AD (i.e. $p(LearnedAC|TargetBC)p(TargetBC) + p(LearnedAC|TargetAD)p(TargetAD)$), compared to the proportion of learners who incorrectly learned the BC or AD pattern, given that their target pattern was AC (i.e. $p(LearnedBC|TargetAC)p(TargetAC) + p(LearnedAD|TargetAC)p(TargetAC)$). The overrepresentation of the gang effect AC pattern observed in Figure 2.5, then, is due to this asymmetry in the direction of mislearning.

Because 5,000 learning steps is enough for the learning agents in this case to reach a relatively stable distribution over pattern types and probabilities, iterating over multiple generations of agents results in little to no change across generations. For this reason, the iterated learning model simulations which iterated over multiple generations limited the amount of data points the learning agents received; here, to 1,000 learning steps, which, in the single-generation simulations, marked approximately the midway point between the beginning of learning and the point at which the distribution over patterns stabilized. For both the random initialization condition and the zero-weight initialization condition, the multi-generation simulations showed a decrease in the proportion of agents learning the gang effect AC pattern across generations. This decrease manifested differently between the initialization conditions. In the zero-weight initialization condition, the proportion of runs in the AC

[4] Because the values in these tables are rounded to two decimal places, there is some noise in the learning of the 0.9-1.0 target patterns which is not visible here. However, the trend still holds for these cases.

Table 2.7. Transition probabilities between pattern types in the simple test typology, divided by target average highest probability range, at 5,000 learning steps for one generation of the iterated learning model, with the zero-weight initialization condition

			Learned		
			BC	AC	AD
Target	0.5-0.6	BC	0.53	0.27	0.20
		AC	0.26	0.42	0.31
		AD	0.14	0.35	0.51
	0.6-0.7	BC	0.69	0.27	0.04
		AC	0.15	0.66	0.19
		AD	0.01	0.34	0.64
	0.7-0.8	BC	0.74	0.26	0
		AC	0.18	0.71	0.11
		AD	0	0.28	0.72
	0.8-0.9	BC	0.97	0.03	0
		AC	0.03	0.95	0.02
		AD	0	0.06	0.94
	0.9-1.0	BC	1	0	0
		AC	0	1	0
		AD	0	0	1

		Target → Learned	
		{BC,AD} → AC	AC → {BC,AD}
Target	0.5-0.6	0.21	0.19
	0.6-0.7	0.20	0.11
	0.7-0.8	0.18	0.10
	0.8-0.9	0.03	0.02
	0.9-1.0	0.00	0.00

pattern begins to decrease immediately, and represents about 3% of runs at the end of the simulation. In the random-weight initialization condition, the proportion of runs in the AC pattern actually increases at first, then begins to decrease, representing around 21% of runs at the end of the simulation. These results can be better understood by examining the transition probabilities at 1,000 learning steps for each initialization condition.

In Table 2.8, the left-hand table shows the transition probabilities at 1,000 learning steps for the zero-weight initialization condition, and the right-hand table shows the proportion of learners who mislearn the BC and AD patterns as AC, compared to the proportion of learners who mislearn the AC pattern as BC or AD. As can be observed in these tables, from target grammars in the 0.5-0.6, the same number of learner agents move into the AC region as move out of it, and from target grammars in the 0.6-0.7 average probability ranges, more learner agents are moving into the region corresponding to the gang effect AC pattern than are moving out of it. On the other hand, from target grammars in the 0.7-0.8 and 0.8-0.9 average probability ranges, more learner agents are moving out of the AC pattern region than are moving into it. The overall decrease in the proportion of runs corresponding to the gang effect AC pattern observed when iterating across generation results (see the left-hand panel in Figure 2.7) is the result of these asymmetries in the direction of mislearning, combined with the trend towards more deterministic grammars across generations. As the target grammars become more deterministic across generations, this increases the number of agents who are learning under simulation conditions in which the likelihood of moving out of the AC region is greater than the likelihood of moving into it.

Table 2.8. Transition probabilities between pattern types in the simple test typology, divided by target average highest probability range, at 1,000 learning steps for one generation of the iterated learning model, with the zero-weight initialization condition

			Learned		
			BC	AC	AD
Target	0.5-0.6	BC	0.53	0.26	0.20
		AC	0.29	0.39	0.32
		AD	0.14	0.34	0.52
	0.6-0.7	BC	0.73	0.23	0.04
		AC	0.20	0.58	0.22
		AD	0.02	0.31	0.67
	0.7-0.8	BC	0.85	0.15	0
		AC	0.29	0.56	0.14
		AD	0	0.19	0.81
	0.8-0.9	BC	0.99	0.01	0
		AC	0.29	0.56	0.14
		AD	0	0.19	0.81
	0.9-1.0	BC	1	0	0
		AC	0	1	0
		AD	0	0	1

		Target → Learned	
		{BC,AD} → AC	AC → {BC,AD}
Target	0.5-0.6	0.20	0.20
	0.6-0.7	0.18	0.14
	0.7-0.8	0.11	0.14
	0.8-0.9	0.07	0.14
	0.9-1.0	0.00	0.00

Table 2.9 shows the transition probabilities, in the left-hand table, and in the right-hand table shows the proportion of learners who mislearn the BC and AD patterns as AC, compared to the proportion of learners who mislearn the AC pattern as BC or AD for the random-weight initialization condition with the iterated learning

model. In this condition, more agents are mislearning the BC and AD patterns as AC than are mislearning the AC pattern as BC or AD, for all but the most deterministic target grammars. This leads to the initial increase in the proportion of runs corresponding to the gang effect AC pattern seen in the right-hand graph in Figure 2.7. The subsequent decrease in the proportion of the AC pattern is due to an interaction with the average probabilities in the target grammar, as discussed with the zero-weight condition. For both initialization conditions, the average probability assigned to the highest probability output candidates increased across subsequent generations of agents. This means that, in each subsequent generation of learning agents, there are more agents tasked with learning a target grammar in the most deterministic group. This in turn means that in each subsequent generation, there are more agents who are more likely to move out of the AC pattern than move into it. The slower, more gradual decrease in the proportion of runs corresponding to the AC pattern in the random-weight initialization condition, compared to the zero-weight condition, is due to the much stronger asymmetry towards mislearning into the AC pattern for less deterministic target grammars.

Table 2.9. Transition probabilities between pattern types in the simple test typology, divided by target average highest probability range, at 1,000 learning steps for one generation of the iterated learning model, with the random-weight initialization condition

			Learned		
			BC	AC	AD
Target	0.5-0.6	BC	0.24	0.63	0.13
		AC	0.09	0.69	0.22
		AD	0.04	0.54	0.42
	0.6-0.7	BC	0.35	0.63	0.02
		AC	0.06	0.79	0.15
		AD	0	0.45	0.55
	0.7-0.8	BC	0.68	0.32	0
		AC	0.17	0.73	0.10
		AD	0	0.29	0.71
	0.8-0.9	BC	0.96	0.04	0
		AC	0.03	0.95	0.02
		AD	0	0.06	0.94
	0.9-1.0	BC	1	0	0
		AC	0	1	0
		AD	0	0	1

		Target → Learned	
		{BC,AD} → AC	AC → {BC,AD}
Target	0.5-0.6	0.39	0.10
	0.6-0.7	0.36	0.07
	0.7-0.8	0.20	0.09
	0.8-0.9	0.03	0.02
	0.9-1.0	0.00007	0.0005

2.4 Discussion

In this chapter, I reported the results of testing both the interactive learning model and the iterated learning model to a simple, hypothetical typological space, in order to investigate any biases that arise from these learning models compared to a sampled baseline distribution. Both models display (1) a robust bias away from the gang effect AC pattern, where agents learn the AC pattern less often than would be expected by chance, and (2) a tendency towards more deterministic grammars, where agents tend to accumulate probability on one output candidate over its competitor. The primary difference between the models lies in the underlying mechanisms that drive their behavior.

As I discussed in §2.2, in the results of the interactive learning model the bias away from the gang effect AC pattern is linked to the bias away from more variable grammars: because the weight pairs which yield the AC pattern are more likely to also yield a more variable pattern, as compared to weight pairs yielding the BC or AD patterns, the agents in the interactive learning model become more likely to learn the BC or AD patterns as they drift towards more deterministic grammars.

In §2.3, I showed how the results for the iterated learning model are dependent upon the number of learning steps allowed to each agent; learning must be cut off before the agents are able to achieve a close fit to their target grammar in order for any change to occur across generations. Within one generation of learning, between a teacher agent and a learner agent, the gang effect AC pattern tends to act as an attractor state for all but the most deterministic grammars. However, this effect is countered by the trend towards more deterministic grammars across generations: as the proportion of agents learning a more deterministic grammar increases, the strength of the AC pattern as an attractor state gets weaker.

For both the interactive and iterated learning models, then, the bias away from the gang effect AC pattern seems to be intertwined with the bias away from variable grammars. In the interactive learning model, agents tend to stay longer in more deterministic grammar states where they are less likely to disagree with each other and trigger an update to the grammar, and in the iterated learning model, learning agents are more likely to mislearn their target grammar in the direction of more deterministic patterns, both of which result in an avoidance of the gang effect AC pattern. The connectedness between these biases is supported by the correlation between the AC pattern and variable grammars in the distribution of the possible patterns across the space of possible weight pairs: because most of the weight pairs which yield the AC pattern lie in more variable regions, avoiding variable grammars leads to a decrease in learning of the AC pattern, and avoiding the AC pattern leads to a decrease in learning of variable patterns.

In the chapters that follow, I will turn to the question of whether either of these biases - away from cumulative constraint interactions and away from variation - are observed in empirical natural language typologies. For both biases, it will be difficult to establish a definitive answer. In the case of cumulative constraint interactions, the difficulty lies in identifying gang effects in natural language patterns. Any given phonological pattern can potentially be analyzed in different ways, in different theoretical frameworks. Patterns which can be analyzed as gang effects in a weighted-constraint grammar could be handled in a ranked-constraint framework if the right constraint set could be cultivated, and so the tradeoff becomes either allowing gang effect patterns, or proliferating the number of constraints. Even assuming a weighted-constraint grammatical theory, the set of possible gang effect patterns depends upon the assumed constraint set. Despite these difficulties, it is possible to use the methodology demonstrated in this chapter to make typological predictions about gang effect patterns, given an assumed underlying grammatical theory. In Chapter 3, I investigate the predictions of the interactive and iterated learning models for a typology of contrast patterns between [s] and [ʃ], which stands in as a specific instance of a general type of typology involving three basic types of constraints: (1) a context-free markedness constraint (here, against [ʃ]), (2) a context-sensitive markedness constraint (here, against [s] followed by a front vowel), and (3) a faithfulness constraint (here, to the feature differentiating [s] and [ʃ]). I find that the gang effect pattern in this typology is indeed vastly underrepresented in the empirical typology, more so than expected by chance. The bias against gang effect patterns generated by the interactive and iterated learning models lines up with needed balance between allowing for their existence, but predicting them to be rarer than expected by chance.

For the case of the bias away from variation, the difficulty lies in the unavailability of comparable empirical typological data. Although various sources exist which compile cross-linguistic typological surveys of observed pattern types or linguistic features, for example the World Atlas of Linguistic Structures (WALS; Dryer & Haspelmath, 2013) and P-base (Mielke, 2008), no database exists in which it is possible to calculate the relative frequencies of deterministic patterns, compared to patterns of variation and/or exceptionality. Compiling such a database lies outside the scope of this work, though I discuss future plans for doing so in Chapter 5. However, although there is a lack of cross-linguistic typological data, there is some experimental literature that investigates how human participants learn from training data exhibiting variation (Reali & Griffiths, 2009; Hudson Kam & Newport, 2005, 2009; Smith & Wonnacott, 2010; Wonnacott, 2011; Hudson Kam, 2015; Samara et al., 2017, among others). In Chapter 4, I further explore the predictions of the iterated and interactive learning models for the relative frequencies of deterministic patterns, variation, and exceptionality, using a typology of variation taken from the artificial grammar learning experiment in Smith & Wonnacott (2010), and compare those predictions to the results observed in the experimental literature.

CHAPTER 3

GANG EFFECT BIAS: THE TYPOLOGY OF CONTRAST TYPES

Chapter 2 demonstrated the application of the interactive and iterated learning models using a simple typological space as a minimal working example, which allowed for fairly straightforward analysis of the models' effects on predicting gradient typological distributions. The results of both the interactive learning model simulations (described in §2.2) and the iterated learning model simulations (described in §2.3) showed a bias against gang effect patterns and a bias away from variation in the grammar. In this chapter, I further investigate the models' bias against gang effect patterns using a typology of possible contrast patterns, and compare the models' predictions against an estimate of the empirically observed typological distribution. Following Carroll (2012), I use as a test case the typology of contrast patterns between [s] and [ʃ], which includes a gang effect pattern - the Elsewhere Neutralization pattern - that Carroll argues is (or was historically) empirically attested in Gujarati. I show how both the iterated and interactive learning models produce a bias away from the gang effect Elsewhere Neutralization pattern, which aligns with the observed rarity of this pattern in the empirical typology.

3.1 Introduction

The test case used throughout this chapter is the typology of contrast patterns between [s] and [ʃ] (see e.g. Carroll, 2012), ranging from full contrast in all environments to no contrast in any environment. In the general case, a typology of contrast patterns is generated through the interactions between three constraints: a general context-free markedness constraint violated by every instance of some feature [αF] in the output, a context-specific markedness constraint violated by every instance of [αF] or [βF] in a specific context, and a faithfulness constraint that penalizes changes in [±F] between the input and the output. The constraints used for the specific case of contrast patterns between [s] and [ʃ] are defined in Table 3.1: a general markedness constraint No[ʃ], a context-specific markedness constraint No[si], and a faithfulness constraint Ident which penalizes changing between [s] and [ʃ].

Table 3.1. Constraints used to model the typology of contrast types

No[ʃ]: (General markedness) Assign one violation for every [ʃ] in the output.
No[si]: (Specific markedness) Assign one violation for every [s] preceding a high front vowel [i] in the output.
Ident: (Faithfulness) Assign one violation for every fricative in the output which disagrees in place with its input correspondent.

In order to focus on the essential predictions generated by the models, I restrict the set of inputs to four possible input forms - /si/, /ʃi/, /sa/, /ʃa/ - which combine the two relevant segments, /s/ and /ʃ/, in the two relevant environments: before high vowels, represented by /i/, and before non-high vowels, represented by /a/. There are two possible outputs for each input: one with [s], and one with [ʃ]. Changes in the vowel are not permitted in this case. The inputs, outputs, and their violation profiles are given in Tableau 3.1.

Tableau 3.1. Tableaux showing the contrast types grammar

/sa/	No[ʃ]	No[si]	Ident
sa			
ʃa	-1		-1

/si/	No[ʃ]	No[si]	Ident
si		-1	
ʃi	-1		-1

/ʃa/	No[ʃ]	No[si]	Ident
sa			-1
ʃa	-1		

/ʃi/	No[ʃ]	No[si]	Ident
si		-1	-1
ʃi	-1		

There are five possible patterns in this typological space, including four which are possible whether the constraints are ranked or weighted, and one, the Elsewhere Neutralization pattern, which is the result of a gang effect and thus only possible when the constraints are weighted. Table 3.2 and Table 3.3 show summarized information about these five patterns: Table 3.2 gives brief descriptions of each of the five patterns, and Table 3.3 gives the weighting conditions for each pattern, as well as the corresponding output mapping for each input form.

Table 3.2. The set of possible patterns in the contrast types typology

(TN) **Total Neutralization**
No contrast between /s/ and /ʃ/; only [s] surfaces

(FC) **Full Contrast**
Contrast between /s/ and /ʃ/ in all environments

(CD) **Complementary Distribution**
[ʃ] occurs before high vowels, and [s] occurs elsewhere

(CN) **Contextual Neutralization**
Contrast between /s/ and /ʃ/ neutralized to [ʃ] before high vowels,
but preserved elsewhere

(EN) **Elsewhere Neutralization**
Contrast between /s/ and /ʃ/ preserved before high vowels,
but neutralized to [s] elsewhere

Table 3.3. Weighting conditions and output forms for the contrast types typology. $w(G)$ means the weight of the general markedness constraint, $w(S)$ means the weight of the context-specific markedness constraint, and $w(F)$ means the weight of the faithfulness constraint.

Pattern	/sa/	/ʃa/	/si/	/ʃi/	Weighting Condition
TN	[sa]	[sa]	[si]	[si]	$w(G) > w(F) + w(S)$
FC	[sa]	[ʃa]	[si]	[ʃi]	$w(F) > w(G) > w(S) - w(F)$
CD	[sa]	[sa]	[ʃi]	[ʃi]	$w(S) - w(F) > w(G) > w(F)$
CN	[sa]	[ʃa]	[ʃi]	[ʃi]	$w(S) - w(G) > w(F) > w(G)$
EN	[sa]	[sa]	[si]	[ʃi]	$w(G) > w(F) > \lvert w(S) - w(G) \rvert$

In the Total Neutralization (TN) pattern, illustrated in Tablau 3.2, there is no contrast between /s/ and /ʃ/; both map to [s] in the output. This pattern occurs when the weight of the general markedness constraint penalizing [ʃ] is greater than the combined weights of the context-specific markedness and the faithfulness constraints.

Tableau 3.2. Tableaux showing the Total Neutralization pattern

	5	1	1	
/sa/	No[ʃ]	No[si]	Ident	*H*
☞sa				0
ʃa	-1		-1	-6
/ʃa/	No[ʃ]	No[si]	Ident	*H*
☞sa			-1	-1
ʃa	-1			-5

	5	1	1	
/si/	No[ʃ]	No[si]	Ident	*H*
☞si		-1		-1
ʃi	-1		-1	-6
/ʃi/	No[ʃ]	No[si]	Ident	*H*
☞si		-1	-1	-2
ʃi	-1			-5

In the Full Contrast (FC) pattern, illustrated in Tableau 3.3, /s/ and /ʃ/ contrast in all environments; both map faithfully in the output. This pattern occurs when the weight of the faithfulness constraint is greater than the weight of the general markedness constraint, allowing /ʃ/ to map faithfully, and when the weight of the context-specific markedness constraint penalizing the sequence [si] is not high enough above either other constraint to prevent /s/ from mapping faithfully before high vowels.

Tableau 3.3. Tableaux showing the Full Contrast pattern

	1	2	4	
/sa/	No[ʃ]	No[si]	Ident	*H*
☞sa				0
ʃa	-1		-1	-5
/ʃa/	No[ʃ]	No[si]	Ident	*H*
sa			-1	-4
☞ʃa	-1			-1

	1	2	4	
/si/	No[ʃ]	No[si]	Ident	*H*
☞si		-1		-2
ʃi	-1		-1	-5
/ʃi/	No[ʃ]	No[si]	Ident	*H*
si		-1	-1	-6
☞ʃi	-1			-1

In the Complementary Distribution (CD) pattern, illustrated in Tableau 3.4, /s/ and /ʃ/ both occur, but do not contrast; both inputs map to output [s] before non-high vowels, and to output [ʃ] before high vowels. This pattern occurs when the weight of the general markedness constraint is higher than the weight of the faithfulness constraint, but the weight of the context-specific markedness constraint is high enough above both other constraints: this blocks /s/ from mapping faithfully before high vowels, so it maps to [ʃ], but blocks /ʃ/ from mapping faithfully before non-high vowels, so it maps to [s].

Tableau 3.4. Tableaux showing the Complementary Distribution pattern

	4	8	1	
/sa/	No[ʃ]	No[si]	Ident	*H*
☞sa				0
ʃa	-1		-1	-5
/ʃa/	No[ʃ]	No[si]	Ident	*H*
☞sa			-1	-1
ʃa	-1			-4

	4	8	1	
/si/	No[ʃ]	No[si]	Ident	*H*
si		-1		-8
☞ʃi	-1		-1	-5
/ʃi/	No[ʃ]	No[si]	Ident	*H*
si		-1	-1	-9
☞ʃi	-1			-4

In the Contextual Neutralization (CN) pattern, illustrated in Tableau 3.5, /s/ and /ʃ/ contrast before non-high vowels, but this contrast is neutralized to [ʃ] before high vowels. This pattern occurs when the weight of the faithfulness constraint is higher than the weight of the general markedness constraint, but the weight of the context-specific markedness constraint is high enough above both other constraints: this allows both /s/ and /ʃ/ to map faithfully before non-high vowels, but blocks /s/ from mapping faithfully before high vowels, so it maps to [ʃ].

Tableau 3.5. Tableaux showing the Contextual Neutralization pattern

	1	8	4	
/sa/	No[ʃ]	No[si]	Ident	*H*
☞sa				0
ʃa	-1		-1	-5
/ʃa/	No[ʃ]	No[si]	Ident	*H*
sa			-1	-4
☞ʃa	-1			-1

	1	8	4	
/si/	No[ʃ]	No[si]	Ident	*H*
si		-1		-8
☞ʃi	-1		-1	-5
/ʃi/	No[ʃ]	No[si]	Ident	*H*
si		-1	-1	-12
☞ʃi	-1			-1

The fifth pattern, Elsewhere Neutralization (EN), is the result of a gang effect between markedness and faithfulness, as illustrated in Tableau 3.6. In the EN pattern, /s/ and /ʃ/ contrast before high vowels, but this contrast is neutralized in all other contexts. EN occurs when the weight of the general markedness constraint is greater than the weight of the faithfulness constraint, and when the weights of both the general and context-specific markedness constraints are sufficiently close together (less than $w(F)$ apart). When the general markedness constraint has a higher weight than the faithfulness constraint, /ʃ/ is blocked from surfacing faithfully before non-high vowels, producing neutralization in the general, “elsewhere” context. This can be most clearly seen in Tableau 3.6(2.), where input /ʃa/ maps to output [sa] with higher probability than to the faithful output [ʃa], which preserves the marked segment. In order for the contrast to emerge in the specific context, before high vowels, the weights of the markedness constraints must be sufficiently close together. This allows the general markedness constraint and the faithfulness constraint to gang up on the context-specific markedness constraint in Tableau 3.6(3.), producing the faithful mapping /si/→[si], and allows the context-specific markedness constraint and the faithfulness constraint to gang up on the general markedness constraint in Tableau 3.6(4.), producing the faithful mapping /ʃi/→[ʃi].

Tableau 3.6. Tableaux showing an example of the Elsewhere Neutralization pattern

1.

	7	6	4		
/sa/	No[ʃ]	No[si]	Ident	*H*	*p*
☞sa				0	1.00
ʃa	-1		-1	-11	0.00

2.

	7	6	4		
/ʃa/	No[ʃ]	No[si]	Ident	*H*	*p*
☞sa			-1	-4	0.95
ʃa	-1			-7	0.05

3.

	7	6	4		
/si/	No[ʃ]	No[si]	Ident	*H*	*p*
☞si		-1		-6	0.99
ʃi	-1		-1	-11	0.01

4.

	7	6	4		
/ʃi/	No[ʃ]	No[si]	Ident	*H*	*p*
si		-1	-1	-10	0.05
☞ʃi	-1			-7	0.95

Carroll (2012) argues that the Elsewhere Neutralization pattern is (or was historically) attested in Gujarati, and that this single instance of attestation means we must not only include this pattern in the typology of possible patterns, but also account for its relative rarity. In a weighted-constraint framework, the Elsewhere Neutralization pattern is already predicted, arising as a gang effect with the same constraint set used to account for the other four patterns in this typology. In this section, I show how the emergent bias against gang effect patterns observed in the interactive and iterated learning models in Chapter 2 greatly reduces the predicted probability of Elsewhere Neutralization.

For the purposes of evaluating the models' predictions against empirical data, I take as a starting point the observed typological counts and frequencies reported by Carroll (2012), shown in Table 3.4. In order to estimate the observed frequency of each of the contrast types in Table 3.3 in natural languages, Carroll calculated the frequency of each type in P-base (Mielke, 2008), a database which includes segment inventories and sound patterns for over 500 languages. The data on which he bases his observed frequency counts represent a bit more than one-third of P-base. According to this estimation, the most common patterns are Total Neutralization then Full Contrast, followed by Complementary Distribution, then Contextual Neutralization. The gang effect pattern, Elsewhere Neutralization, falls below all the others, representing an estimated 0.5% of the typology. There is no available data from which to estimate the empirical distribution across average highest probabilities.

Because these observed frequencies are estimated over a sample of P-base, which is itself a sample of known natural languages, the exact frequency values should not be taken at face value. In order to test the robustness of the estimated empirical distribution, I used bootstrap sampling to establish 95% confidence intervals over the reported frequency of each pattern in the typology, shown in Table 3.4, over both counts and proportions. Using the raw counts reported by Carroll (2012) as the base data set, and sampling with replacement, I took 10,000 samples of 192 patterns. Then, for each pattern, I listed the number of times that pattern occurred in each sample, and took the 2.5% percentile and the 97.5% percentile as the upper and lower bound for the 95% confidence interval.

Table 3.4. Estimated empirical frequencies and 95% confidence intervals for each pattern type in the contrast types typology

Pattern	Raw Count	Proportion	95% CI (Count)	95% CI (Prop)
TN	81	0.44	67 - 94	0.36 - 0.51
FC	68	0.37	55 - 81	0.30 - 0.44
CD	19	0.10	11 - 28	0.06 - 0.15
CN	15	0.08	8 - 23	0.04 - 0.13
EN	1	0.005	0 - 3	0.00 - 0.02
Total	192			

In analyzing these confidence intervals, it can be seen that a total ordering between the five patterns in the contrast type typology cannot be reliably established. The confidence intervals for the Total Neutralization and Full Contrast patterns overlap substantially, as do those for the Complementary Distribution and Contextual Neutralization patterns. We can, however, reliably identify three tiers of estimated frequency: Total Neutralization and Full Contrast are reliably more frequent than Complementary Distribution and Contextual Neutralization, and the Elsewhere Neutralization pattern is reliably the least frequent.

Before evaluating the effects of the interactive and iterated learning models on this typology, and evaluating them against the estimated empirical distribution, it is necessary to first establish a baseline gradient typological prediction using the grammatical space alone. The baseline prediction over pattern types was generated by randomly sampling weights for each of the constraints, and then categorizing the resulting pattern by identifying the output candidate assigned highest probability, for each input. The prediction over average highest probabilities was generated by averaging together the probabilities assigned to the highest probability output candidates for each input. Weights were sampled from a uniform distribution with the range 0-10, and 25,000 sets of constraint weights were sampled. The predicted distributions are given in Table 3.5, along with the empirical distribution again for comparison. In this baseline, the largest proportion of the space of possible weights corresponds to the Full Contrast pattern (0.42), followed surprisingly by the gang effect Elsewhere Neutralization pattern (0.25), which has the lowest empirical frequency, and then Total Neutralization (0.17), which has the highest empirical frequency. The Complementary Distribution and Contextual Neutralization patterns have the lowest frequency in the sampled baseline distribution (each 0.08). Additionally, the majority of samples taken in the baseline correspond to a pattern in the highest range of average probabilities, with the proportions of patterns decreasing as the range of average probabilities decreases.

Table 3.5. The sampled baseline distribution over pattern types and average highest probabilities for the contrast types typology

	Pattern					Average Highest Probability				
	TN	FC	CD	CN	EN	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Empirical	0.44	0.37	0.10	0.08	0.005	—	—	—	—	—
Sampled Baseline	0.17	0.42	0.08	0.08	0.25	0	0.001	0.03	0.25	0.72

In using the observed typological frequencies estimated by Carroll (2012) to evaluate the performance of the interactive and iterated learning models, there are two major considerations to take into account. The first is that these frequencies represent an estimation over a sample of a sample of known natural languages. I have attempted to mitigate any potentially spurious influences from this estimation by using bootstrap sampling to establish 95% confidence intervals over the empirical distribution, as discussed above (see Table 3.4). Based on these confidence intervals, rather than risk overfitting the models to the exact estimated values, I evaluate the models against the three reliable tiers of frequency, which I mark in Table 3.5 and in later tables with the dashed lines.

The second important consideration is that the biases that emerge in these learning models constitute a kind of structural, or analytic, bias for grammars which can be learned more easily or more quickly. The learning agents in these models have no components corresponding to articulatory production or auditory perception, or any kind of channel bias, which are factors known to influence language learning in humans (see Moreton & Pater, 2012a,b, for further discussion). These factors likely have an additional influence on the typology of contrast types which is not accounted for in the interactive and iterated learning models as implemented in this work, as I discuss further in §3.4.

3.1.1 Visualizing the contrast type typology

For the simple typology in Chapter 2, the space of possible weights and grammars could be visualized as regions in a two-dimensional plane, where each dimension corresponded to the weight of one of the two constraints in the typology. For the typology of contrast types, the space of possible weights and grammars can be visualized as regions in a three-dimensional plane, where each dimension corresponds to one of the three constraints in this typology: No[ʃ] (GEN), No[si] (SPEC), and Ident (FAITH).

Figure 3.1 and Figure 3.2 plot the typology of contrast types in a three-dimensional cube, where the upper bound on constraint weights was arbitrarily chosen as 10. Figure 3.1 shows an off-center view from the origin, which is the bottom center corner. Figure 3.2 shows a view from the right of the origin, where the origin is at the bottom left. Each figure plots the planes that mark the borders between pattern types, each of which is composed of all sets of weights which generate equal probability between the competing output candidates for some input. The green plane plots all the weights which give equal probability to the mappings /si/→[si] and /si/→[ʃi]. The CD and CN patterns, which map /si/ to [ʃi], lie on one side of this plane, while the FC, TN, and EN patterns, which map /si/ to [si], lie on the other. The orange plane plots all the weights which give equal probability to the mappings /ʃa/→[ʃa] and /ʃa/→[sa]. The FC and CN patterns, which map /ʃa/ to [ʃa], lie on one side of this plane, while the TN, CD, and EN patterns, which map /ʃa/ to [sa], lie on the other. The purple plane plots all the weights which give equal probability to the mappings /ʃi/→[si] and /ʃi/→[ʃi]. The TN pattern, the only one which maps /ʃi/ to [si], lies on one side of this plane, while the FC, CD, CN, and EN patterns, which map /ʃi/ to [ʃi], lie on the other. The gang effect EN pattern, like the gang effect AC pattern in Chapter 2, is surrounded on all sides by other patterns; it is the only one which does not share a face with an axis. The relative volumes of the regions corresponding to each pattern type correspond to the sampled baseline probabilities shown in Table 3.5.

Figure 3.1. Plot of the contrast types typology space - view from origin

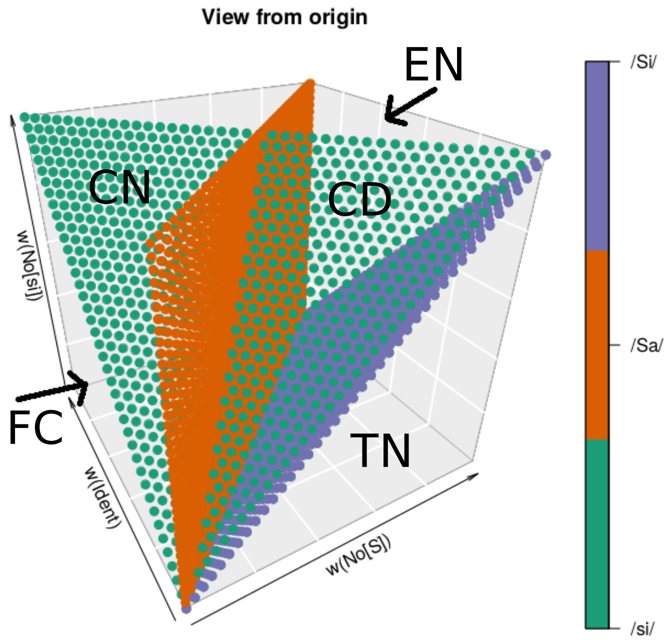
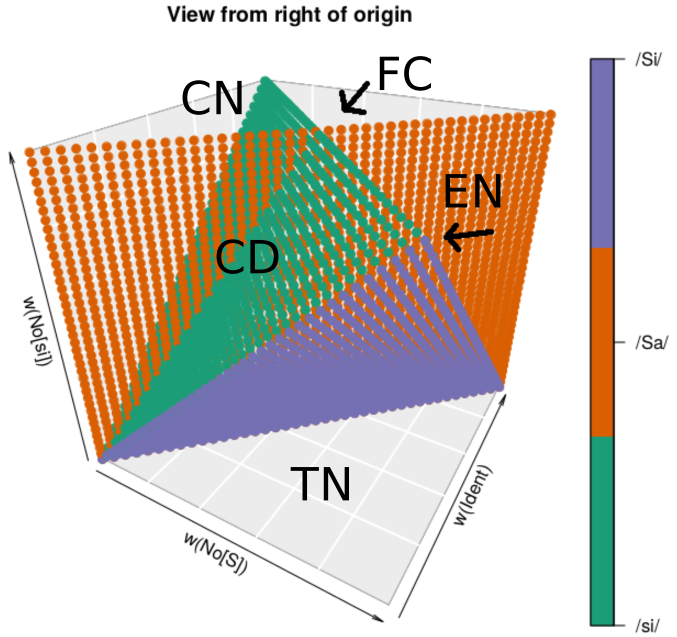


Figure 3.2. Plot of the contrast types typology space - view from right of origin



In the remainder of this chapter, I will discuss first the results of applying the interactive learning model to this typological space in §3.2, then in §3.3 I discuss the results of applying the iterated learning model. In the results of both models, the Total Neutralization and Full Contrast patterns are predicted to be the most frequent, as observed in the estimated empirical frequencies in Table 3.5, but the predicted frequency of the gang effect Elsewhere Neutralization pattern is not substantially different from the predictions for Complementary Distribution or Contextual Neutralization, and in some results with the iterated learning model is actually higher. However, in all cases, the Elsewhere Neutralization pattern is predicted to be substantially less frequent than would be expected by chance, given the baseline probabilities in Table 3.5. §3.4 provides an overall discussion.

3.2 Interactive learning simulations

In this section, I present the results of applying the interactive learning model to this typology of contrast patterns between /s/ and /ʃ/. As with the simulations reported for the simple typology in Chapter 2, I test two initialization conditions: one in which both agents are initialized with constraint weights at zero (the zero-weight initialization condition), and one in which both agents are initialized with random constraint weights (the random-weight initialization condition).

In the zero-weight initialization condition, each learning agent was initialized with all constraint weights at zero. This gives the learning agents an unbiased starting point; with all weights at zero, the set of competing output candidates for each input have equal probability ($p = 0.5$). 10,000 runs with this initialization were performed.

In the random-weight initialization condition, the learning agents were initialized with constraint weights which were randomly sampled from a uniform distribution with range 0.0-10.0, and both agents were assigned the same set of initial weights. Because this case is intended to emulate a naturalistic typology, the sampled initial states were not balanced across pattern type or average highest probability, preserving the prior biases in the typological space. 10,000 runs with this initialization were performed.

For simulations in both initialization conditions, the learning agents exchanged data for 100,000 learning steps, with a learning rate of 0.1.

3.2.1 Simulation results

Table 3.6 summarizes the agents' distribution over pattern types and average highest probabilities at the end of the simulations. The table shows results for both initialization conditions, as well as repeating the sampled baseline distribution. For both initialization conditions, the distribution over pattern types approaches the three-tiered nature of the empirical typology (see Table 3.5). The Total Neutralization and

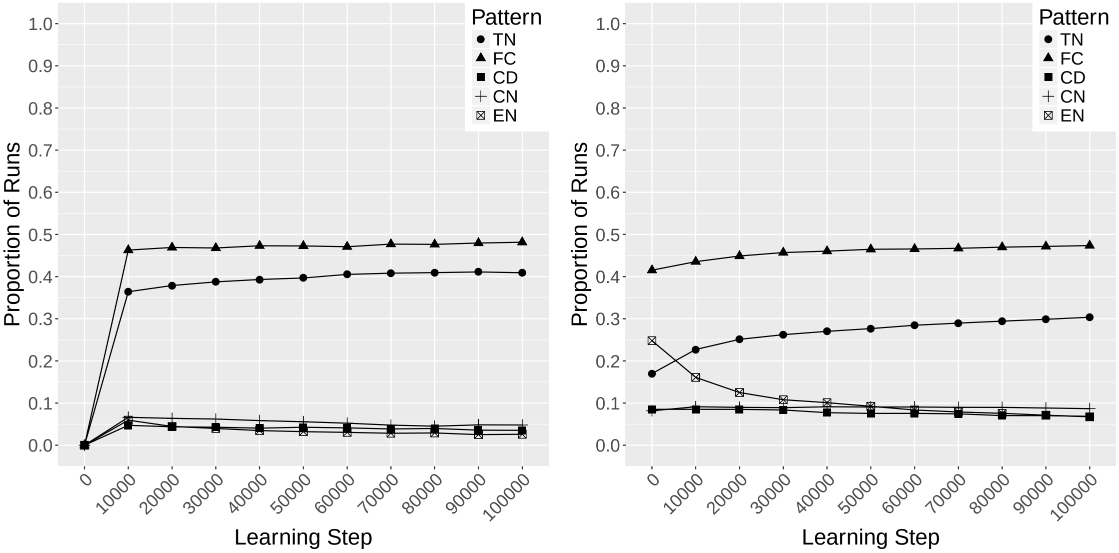
Full Contrast patterns are predicted to be most frequent, followed by Complementary Distribution and Contextual Neutralization, and the gang effect Elsewhere Neutralization pattern. The prediction of rarity for the Elsewhere Neutralization pattern, however, is not the extreme difference that was seen in the empirical typology - there, the EN pattern is estimated to occur much more infrequenly than any other pattern, whereas the predictions from both the random-weight condition and the zero-weight condition differ only minimally from the predictions for CD or CN.

Table 3.6. The distribution over pattern types and average highest probabilities after learning with the interactive learning model in the contrast types typology

	Pattern					Average Highest Probability				
	TN	FC	CD	CN	EN	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.17	0.42	0.08	0.08	0.25	0	0.00	0.03	0.25	0.72
Zero-weight Cond.	0.41	0.48	0.04	0.05	0.03	0	0	0.00	0.02	0.98
Rand-weight Cond.	0.30	0.47	0.07	0.09	0.07	0	0	0.00	0.01	0.99

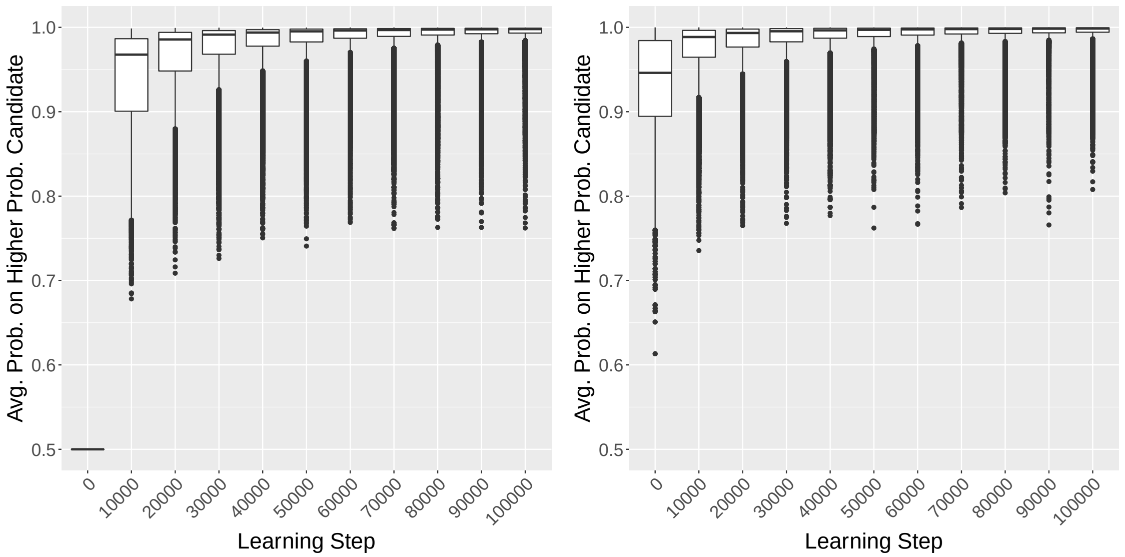
Taking a closer look at the distribution over pattern types, the graphs in Figure 3.3 plot the proportion of runs corresponding to each pattern type across learning steps, tracking the change in the distribution over patterns as the simulation progresses. The left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. For the zero-weight condition, at zero learning steps, the agents are still in the initial state with equal probability on all output candidates, and thus are not categorized as belonging to any pattern. For the random-weight initialization condition, at zero learning steps, the agents are still in the random initial states, where the distribution is the same as in the sampled baseline. In both initialization conditions, the agents diverge fairly quickly into a distribution with two tiers, and the distribution remains fairly stable as the simulation continues, with only relatively minor changes after about 20,000 learning steps. The Full Contrast and Total Neutralization patterns are favored heavily over the other three patterns, but the gang effect Elsewhere Neutralization pattern does not diverge significantly from Complementary Distribution and Contextual Neutralization.

Figure 3.3. Plots showing the change in the distribution over pattern types across learning steps with the interactive learning model in the contrast types typology (Left: zero-weight condition, Right: random-weight condition)



The graphs in Figure 3.4 plot the average probabilities assigned to the highest probability output candidates across learning steps, tracking the change in the distribution over probabilities as the simulation progresses. Again, the left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. As the learning agents exchange data in both conditions, they gradually accumulate a greater majority of probability onto the highest probability candidates, as seen in both the increase in median probability and the decrease in variance across the distributions.

Figure 3.4. Plots showing the change in the distribution over average highest probabilities across learning steps with the interactive learning model in the contrast types typology (Left: zero-weight condition, Right: random-weight condition)



3.2.2 Interactive learning discussion

The above sections describe the results of applying the interactive learning model to the typology of contrast patterns between /s/ and /ʃ/ under two initialization conditions: one in which both agents are initialized with constraint weights at zero, and one in which both agents are initialized with randomly sampled constraint weights. As shown in the results for both initialization conditions, the interactive learning model demonstrates a bias towards the Total Neutralization and Full Contrast patterns, which is in line with the observation that these are the most common patterns in the empirical typology, as well as a bias towards more deterministic patterns. The model does not, however, consistently predict that the gang effect pattern, Elsewhere Neutralization, should be the least common pattern, but instead predicts it to be about as frequent as the Complementary Distribution and Contextual Neutralization patterns.

The advantage for the TN and FC patterns here is linked to the bias towards more deterministic patterns. Recall that, in the interactive learning model results for the simple test typology in Chapter 2, the bias away from the gang effect pattern was found to be linked to the agents' bias away from variation in the grammar. As the agents developed more deterministic grammars, they became less likely to learn the gang effect pattern, because proportionally more of the set of possible weights corresponding to that pattern also produced more variable grammars, when compared to the other patterns.

In the case of the typology of contrast patterns, the advantage for the TN and FC patterns observed in these results stems from the same phenomenon. Where the gang effect pattern in the simple typology had proportionally more variable possible grammars, the TN and FC patterns in the typology of contrast patterns have proportionally fewer variable possible grammars, compared to the other patterns in this typology. This can be seen in Table 3.7, which gives a more detailed presentation of the baseline distribution obtained by random sampling (see Table 3.5). This table shows, over samples corresponding to each pattern type, what proportion of the sampled sets of weights correspond to each range of average highest probability (for example, out of all of the sampled sets of weights which correspond to a TN pattern, 77% of them also correspond to a grammar in the 0.9-1.0 average probability range). While very few sampled grammars fall into the three lowest ranges, it can still be seen that sets of weights which yield grammars in the TN and FC patterns contain a higher proportion of grammars in the highest range (0.77 and 0.83, respectively) compared to the other three patterns (0.53, 0.60, and 0.59), but conversely fewer grammars across the lower ranges of average highest probability. The consequence of this is that, where having more variation put the gang effect pattern at a disadvantage in the simple typology, having less variation gives TN and FC an advantage over the other patterns in the typology of contrast types.

The bias away from variation in the grammar does not yield a disadvantage for the gang effect Elsewhere Neutralization pattern in the typology of contrast types.

Table 3.7. Table showing the proportion of interactive learning model simulations in each range of average highest probabilities, normalized by pattern type, for the typology of constrast types

	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
TN	0.00	0.00	0.02	0.21	0.77
FC	0.00	0.00	0.01	0.15	0.83
CD	0.00	0.00	0.06	0.41	0.53
CN	0.00	0.00	0.03	0.37	0.60
EN	0.00	0.00	0.03	0.38	0.59

This is because, as observed in Table 3.7, the EN pattern contains a very similar proportion of sampled weights in each probability range as the CD and CN patterns. Thus, the bias for deterministic grammars does not, in the end, differentiate between these three patterns.

Despite the lack of difference between the proportions of Elsewhere Neutralization, Complementary Distribution, and Contextual Neutralization in the final distribution over pattern types, the EN pattern is the only one predicted to be substantially less frequent than expected given the baseline distribution. As these results show (see Table 3.6 and Figure 3.3), the predicted frequencies of FC, CD, and CN remain fairly close to their baseline frequencies as learning progresses, while the predicted frequency of TN is substantially increased (0.17 baseline, compared to 0.41 or 0.30 after learning) and the predicted frequency of EN is substantially decreased (0.25 baseline, compared to 0.03 or 0.07 after learning). Thus, although the results of the interactive learning model do not consistently differentiate the gang effect EN pattern from CD and CN, the model does substantially decrease the predicted probability of the gang effect pattern. The further disadvantage for EN observed in the empirical typology likely stems from other biases not accessible in this model, such as articulatory or perceptual biases, as discussed in §3.4.

3.3 Iterated learning simulations

In this section, I present the results of applying the iterated learning model to this typology of contrast patterns between /s/ and /ʃ/. As with the simulations reported for the simple typology in Chapter 2, I again test two initialization conditions for new child agents introduced into the chain: one in which new learner agents are initialized with constraint weights at zero, and one in which new learner agents are initialized with random constraint weights, sampled from a uniform distribution with the range 0.0-10.0. In all cases, the initial teacher agents in each chain have grammars which are randomly sampled, and as with the interactive learning model simulations, the sampled initial states were not balanced across pattern type or average highest probability range, preserving the prior biases in the typological space. The learning rate was 0.1, and 10,000 runs of each simulation were performed.

The results here are presented in two sections, one showing the learning paths of a single generation of agents learning from their fixed target grammars for 5,000 learning steps, and the second showing the results of iterating over multiple generations of agents, who each learn from their target grammar for 1,000 learning steps.

3.3.1 Simulation results: within one generation

This section presents the results of transmitting the balanced set of randomly sampled initial states from the initial teacher to the first generation of learners. Observing the trends in the transmission of a pattern from teacher to learner will aid in understanding the trends observed in the transmission of patterns across multiple generations of agents. In the results presented below, each learner agents received data from its teacher agents for 5,000 learning steps.

Table 3.8 shows the final distributions over pattern types and probability categories across runs for the learner agents in both initialization conditions, as well as giving the sampled baseline distribution for comparison. Because the initial grammars for the teacher agents were randomly sampled, the baseline distribution also corresponds to the learner agents' target distribution. The agents in both the zero-weight and random-weight initialization conditions match the target distribution incredibly closely.

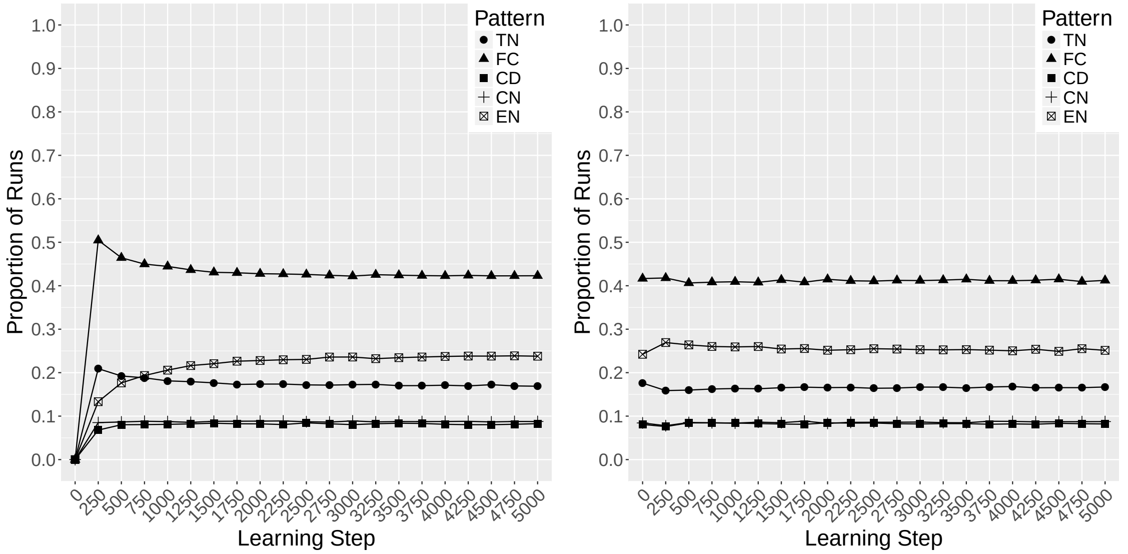
Table 3.8. The distribution over pattern types and average highest probabilities after one generation of learning with the iterated learning model in the contrast types typology

	Pattern					Probability Bin				
	TN	FC	CD	CN	EN	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.17	0.42	0.08	0.08	0.25	0	0.001	0.03	0.25	0.72
Zero-weight Cond.	0.17	0.42	0.08	0.09	0.24	0	0.002	0.02	0.25	0.72
Rand-weight Cond.	0.17	0.41	0.08	0.09	0.25	0	0.002	0.02	0.24	0.74

The graphs in Figure 3.5 take a closer look at the distribution over pattern types, plotting the proportion of runs in which the learner agents hold each pattern type across learning steps. The left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. At zero learning steps, the learner agents are in their initial state; in the zero-weight initialization condition, this means weights at zero and thus no preference for any output forms, and in the random-weight initialization condition, this produces the distribution seen in the sampled baseline. In both conditions, the agents quickly settle into a distribution closely matching the target distribution.

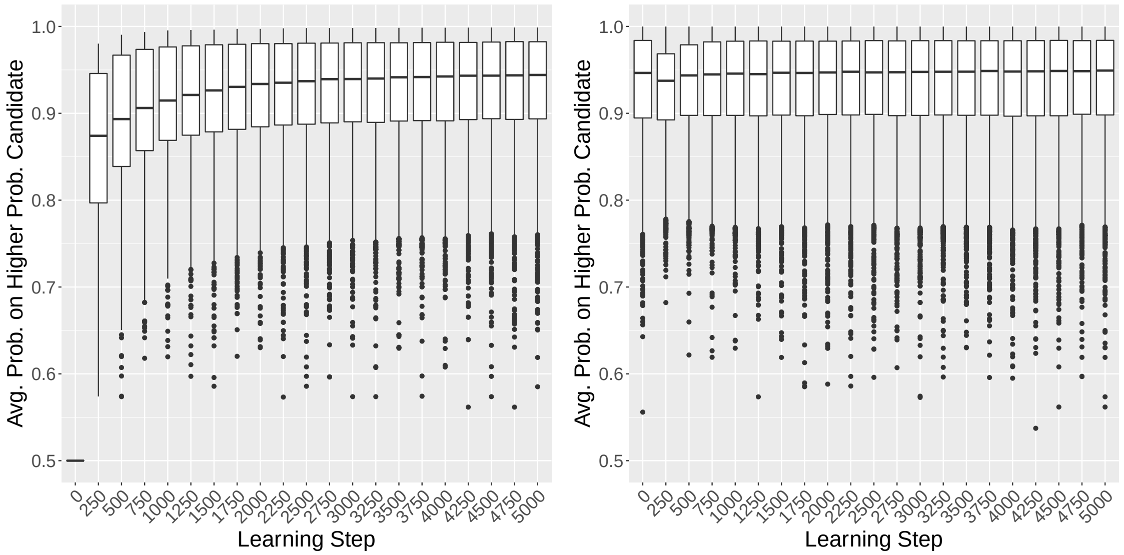
Figure 3.6 takes a closer look at how the average probability the learner agents assign to the highest probability output candidates changes throughout learning, plotting the distribution over average probabilities by learning step. Again, the left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. In the zero-weight initialization condition, the average probabilities assigned by the learning agents increases to match the target distribution; this is because, at zero learning

Figure 3.5. Plots showing the change in the distribution over pattern types across one generation of learning with the iterated learning model in the contrast types typology (Left: zero-weight condition, Right: random-weight condition)



steps, the agents assign equal probability among competing output candidates, and then gradually increase their constraint weights to fit their target grammar. In the random-weight initialization condition, on the other hand, the learner agents' initial distribution is already similar to the target distribution, though for each specific pair the learner and teacher agents' grammars do not necessarily match. The learner agents' movement from their initial distribution to their target distribution deceptively appears to be simply noise in the initial stages of learning.

Figure 3.6. Plots showing the change in the distribution over average highest probabilities within one generation of learning with the iterated learning model in the constrast types typology (Left: zero-weight condition, Right: random-weight condition)



3.3.2 Simulation results: across generations

This section shows the results of iterating learning over 200 generations, where each new learner agent learns from its teacher agent for 1,000 learning steps. As seen above in the presentation of the simulation results within one generation, 5,000 learning steps is enough for the learning agents in this system to reach a relatively stable distribution over pattern types and probabilities. Iterating over multiple generations of agents, when each generation is able to reach this stable point, results in little to no change across generations. In order to have a chance to observe any biases at all emerge from this model, the agents must be cut off at a point earlier in learning. Because the distributions over patterns, seen in Figure 2.5, appear to become relatively stable around 2,000 learning steps, the learning agents in the simulations presented in this section were given 1,000 learning steps: the midway point between the initial state and the stable state.

Table 2.5 shows the distributions over pattern types and probabilities after 200 generations for both initialization conditions, as well as giving the sampled baseline distribution for comparison. In both initialization conditions, the Total Neutralization and Full Contrast patterns are predicted to be more frequent than the other patterns in the typology. In the zero-weight condition, the gang effect Elsewhere Neutralization pattern is predicted to be about as frequent as the Complementary Distribution and Contextual Neutralization Patterns, as was the case in the interactive learning model results in §3.2. In the random-weight initialization condition, however, the predicted probability of the EN pattern is still relatively high, when compared to the other simulation results. The advantage for the TN pattern is weaker in this condition as well.

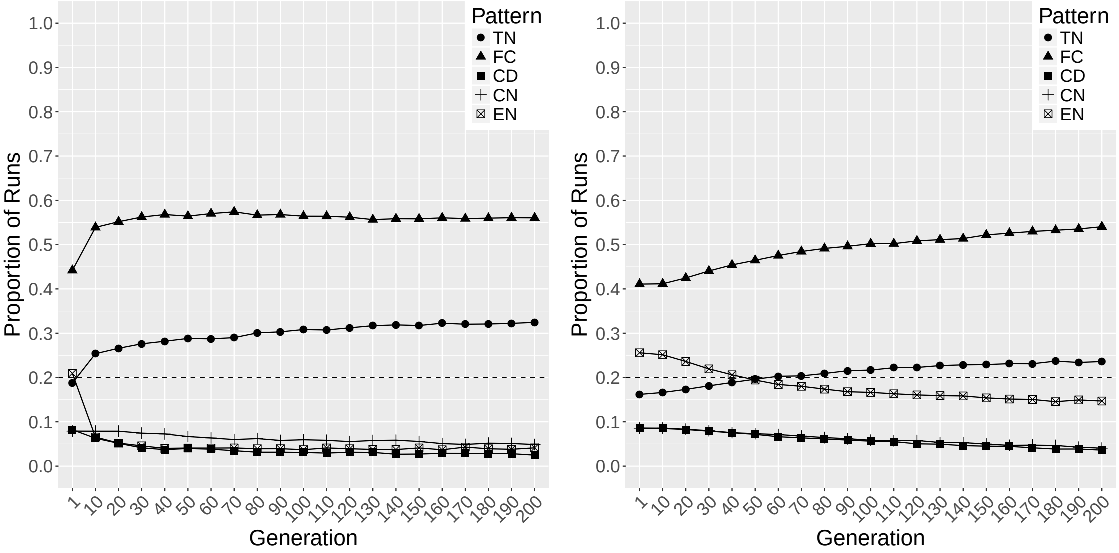
Table 3.9. The distribution over pattern types and average highest probabilities after 200 generations of learning with the iterated learning model in the contrast types typology

	Pattern					Probability Bin				
	TN	FC	CD	CN	EN	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.17	0.42	0.08	0.08	0.25	0	0.001	0.03	0.25	0.72
Zero-weight Cond.	0.32	0.56	0.03	0.05	0.04	0	0.01	0.13	0.23	0.64
Rand-weight Cond.	0.24	0.54	0.04	0.04	0.15	0	0	0	0.09	0.91

The graphs in Figure 3.7 take a closer look at the distribution over pattern types, plotting the proportion of runs in which the learner agents hold each pattern type across learning steps. Again, the left-hand graph shows the results of the zero-weight initialization condition, and the right-hand graph shows the results of the random-weight initialization condition. In the zero-weight condition, the proportion of runs corresponding to each pattern quickly settles into a fairly stable and stratified distribution. In the random-weight condition, the rate of change in the distribution is slower across generations, with the proportion of TN and FC patterns rising and the proportion of CD, CN, and EN patterns slowly falling.

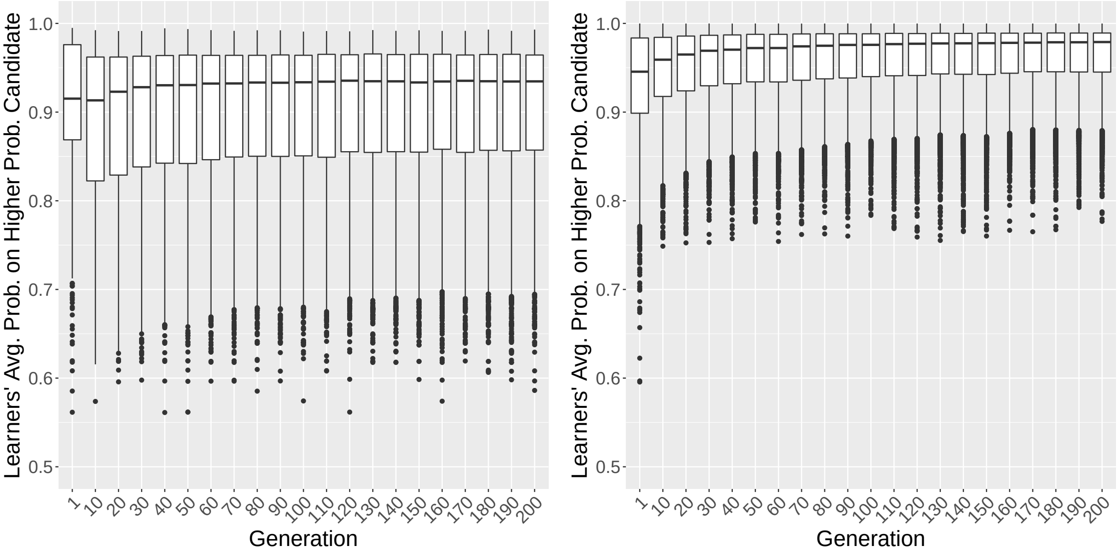
The graphs in Figure 3.8 plot the change in the distribution over the probabilities assigned to higher probability candidates across generations of learner agents (zero-

Figure 3.7. Plots showing the change in the distribution over pattern types across 200 generations of learning with the iterated learning model in the contrast types typology (Left: zero-weight condition, Right: random-weight condition)



weight condition results are on the left, and random-weight condition results are on the right). As generation increases, the mean average probability increases, and the variance of the distribution over probabilities decreases, showing that the bias towards more deterministic grammars is compounded as the grammars are iterated over multiple generations of agents. However, this effect is weaker for the zero-weight initialization condition than for the random-weight initialization condition.

Figure 3.8. Plots showing the change in the distribution over average highest probabilities across 200 generations of learning with the iterated learning model in the contrast types typology (Left: zero-weight condition, Right: random-weight condition)



3.3.3 Iterated learning discussion

The above sections describe the results of applying the iterated learning model to the typology of contrast patterns between /s/ and /ʃ/, looking at both the effects within one generation (i.e. from initial teacher to first learner) and the effects of iterating across a number of generations. Two initialization conditions were considered: one in which new learner agents were initialized with constraint weights of zero, and one in which new learner agents were initialized with constraint weights randomly sampled from a uniform distribution with the range 0.0-10.0. In all cases, the grammars of the first teachers were generated by random sampling constraint weights from a uniform distribution with the range 0.0-10.0.

The results for simulations within one generation, from one teacher agent to the next learner, show that by about 2,000 learning steps, the learning agents settle into a stable distribution over patterns which matches the randomly sampled target distribution. This shows that, given sufficient data, each pattern type in this typology is learnable. In order to see any change emerge when patterns are iterated over multiple generations of agents, the amount of data each learner receives from its teacher must be restricted. Here, each generation of learning agents is restricted to 1,000 learning steps, the halfway point between initialization and (apparent) convergence to the target distribution.

Examining the results of iterated over multiple generations of agents reveals biases similar to those observed in the results of the interactive learning simulations. For both the zero-weight and random-weight initialization conditions, the TN and FC patterns are predicted to be more frequent than the other three patterns (see Table 3.8), which mirrors the trend observed in the empirical distribution in Table 3.5. However, the advantage for the TN and FC patterns is weaker in the random-weight condition than in the zero-weight condition (see Figure 3.5). This is mainly due to the weaker disadvantage for the gang effect EN pattern in the random-weight condition, compared to the zero-weight condition. The results for the zero-weight condition are similar to those observed in the results of the interactive learning model, where the predicted frequency of the EN pattern is not substantially different than that of the CD and CN patterns. In the results of the random-weight condition, on the other hand, the predicted frequency of the EN pattern is still substantially more than the CD or CN patterns. The probability of the EN pattern is only reduced about half as much compared to the baseline (0.25 to 0.15), as it is in the zero-weight condition (0.25 to 0.04).

This difference between the zero-weight and random-weight initialization conditions can be understood by examining the transition probabilities between pattern types at 1,000 learning steps in each condition. These are shown in Table 3.10 for the zero-weight initialization condition, and in Table 3.11 for the random-weight initialization condition. In the zero-weight condition, the Elsewhere Neutralization pattern is less likely to be learned faithfully, and other patterns are less likely to be mislearned as EN, when compared to the random-weight condition. This difference results in the overall slower rate of decrease for the EN pattern in the random-weight condition.

Table 3.10. Transition probabilities between pattern types in the contrast types typology at 1,000 learning steps within one generation of the iterated learning model, in the zero-weight initialization condition

		Learned				
		TN	FC	CD	CN	EN
Initial	TN	**0.69**	0.08	0.04	0.01	0.18
	FC	0.06	**0.71**	0.04	0.10	0.10
	CD	0.04	0.05	**0.63**	0.15	0.13
	CN	0.01	0.14	0.16	**0.66**	0.05
	EN	0.15	0.18	0.10	0.04	**0.53**

Table 3.11. Transition probabilities between pattern types in the contrast types typology at 1,000 learning steps within one generation of the iterated learning model, in the random-weight initialization condition

		Learned				
		TN	FC	CD	CN	EN
Initial	TN	**0.60**	0.05	0.07	0.01	0.27
	FC	0.04	**0.58**	0.07	0.10	0.22
	CD	0.03	0.03	**0.71**	0.11	0.13
	CN	0.01	0.10	0.22	**0.62**	0.05
	EN	0.12	0.09	0.12	0.03	**0.65**

Despite these differences, the iterated learning model still reduces the predicted probability of the gang effect EN pattern to a greater extent, relative to the other patterns in the typology. As will be discussed in §3.4, other biases not accessible in this model, such as articulatory or perceptual biases, likely play a further role in reducing the observed infrequency of the EN pattern.

3.4 Discussion

In this chapter, I reported the results of testing both the interactive learning model and the iterated learning model on the typology of contrast types between /s/ and /ʃ/, and compared these results to an estimate of the empirical typological distribution over pattern types. Rather than trying to fit a full ordering over the five pattern types, I evaluated the models on their ability to predict three reliable tiers of frequency: that Total Neutralization and Full Contrast should be more common than Complementary Distribution and Contextual Neutralization, and that the gang effect Elsewhere Neutralization pattern should be the least common. I additionally explored the models' predictions for patterns of variation, although no empirical estimation for this data is available for comparison.

In §3.2 and §3.3, I show how both the interactive learning model and the iterated learning model predicted that the TN and FC patterns should be the most common, as well as greatly reduced the predicted probability of the gang effect Elsewhere Neutralization pattern, relative to the sampled baseline. However, the EN pattern was not reliably predicted to be more rare than the Complementary Distribution and Contextual Neutralization patterns, as in the estimated empirical typology. For the iterated learning model, the bias away from the gang effect EN pattern was stronger in the zero-weight initialization condition than in the random-weight initialization condition, where the EN pattern receives a higher predicted probability than the CD or CN patterns. The interactive learning model additionally predicts a strong tendency towards more deterministic patterns, as the agents gradually accumulate a majority of probability on one output candidate over it's competitor, whereas this pressure is somewhat weaker in the iterated learning model results.

Although the gang effect EN pattern is not reliably predicted to be the least frequent pattern type in all simulations, both of the models still significantly reduce the predicted probability of the EN pattern relative to the sampled baseline distribution, and relative to the impact on other patterns in the typology. This is a strong result, especially as the learning models used here represent only one kind of bias thought to influence typology: an analytic bias, which favors patterns which are easier or faster to learn given assumptions about the structure of the grammatical space (see Moreton

& Pater, 2012a, for more discussion). The empirical distribution over pattern types in this typology is potentially influenced by other factors in addition to an analytic bias, such as accidents of history and cultural influence, but especially channel and substantive biases, although the evidence supporting the influence of substantive bias on typology is weaker than for the influence of analytic bias (see Moreton & Pater, 2012b).

There are several factors that may contribute to the kinds of channel and substantive biases that would potentially influence the typology of contrast types between /s/ and /ʃ/, including several articulatory and perceptual biases that relate to these segments and their relations to neighboring vowels. One articulatory bias is that the articulation of the palatal fricative [ʃ] is more similar to the high front vowel [i] than articulation of the alveolar fricative [s] is to [i] (Gafos, 1999; Goldstein & Fowler, 2003; Bateman, 2007). This bias would favor neutralizing a contrast between /s/ and /ʃ/ to [ʃ] before [i], thus favoring the Complementary Distribution and Contextual Neutralization patterns, and disfavoring the Full Contrast and Elsewhere Neutralization patterns. Another articulatory bias is that palatal consonants are potentially more difficult to articulate than alveolar consonants (Lindblom et al., 2011; Stevens, 1989), which would favor neutralizing the contrast between /s/ and /ʃ/ at least outside of palatalizing contexts, and so favoring the Total Neutralization and Complementary Distribution patterns, as well as potentially favoring the Elsewhere Neutralization pattern. An example of a perceptual bias is that the formant transitions into the vowel [i] make [s] is more easily confusable with [ʃ] in this context (Stevens, 1989; Boersma & Hamann, 2008; Peterson & Barney, 1952), which would again favor neutralizing the contrast in this context, thus favoring the CD and CN patterns over the FC and EN patterns. These substantive phonetic biases potentially provide an extra advantage for the CD and CN patterns over the gang effect EN pattern, which could help derive the distribution seen in the empirical typology if acting in concert with the strong analytic bias produced by the interactive and iterated learning models.

One potential confound for this reasoning is that the nature of the constraint set used for this typology is constructed to already incorporate these articulatory and perceptual biases into the set of possible grammars that can be represented. The constraint set includes a general markedness constraint against the segment [ʃ] (No[ʃ]), but not one against [s], reflecting the articulatory advantage for [s] over [ʃ]. The constraint set also includes a context-specific markedness constraint against the segment [s] before the high front vowel [i] (No[si]), but not one against [ʃ] before [i], reflecting the articulatory and perceptual advantage for palatals before front vowels. Another potential difficulty for this line of reasoning is that, as stated above, the evidence supporting the influence of substantive biases on typology is weaker than the evidence supporting the influence of analytic biases. Moreton & Pater (2012a,b) provide a survey of previous work, focusing on artificial language learning experiments which compare success in learning artificial patterns differing in structural complexity (Moreton & Pater, 2012a) or phonetic substance (Moreton & Pater, 2012b). This survey reveals that while studies which compare patterns differing in structural complexity consistently find a bias for simpler patterns, the findings from studies which compare patterns differing in phonetic substance have largely produced mixed results or null effects. In order to fully settle the question of how substantive biases affect typological trends, it is first necessary to settle the questions of whether substantive biases do indeed affect typological trends, whether there is an innate universal constraint set shaped by articulatory and perceptual factors, or if the set of constraints is induced from the learning data, and the question of whether substantive biases have an influence on learning outside of, or in addition to, their influence on the set of constraints that learners may use to represent their language. These questions lie outside of the scope of this work.

In the next chapter, I give a more detailed investigation into the predictions for the relative frequencies of deterministic patterns and patterns of exceptionality generated by the interactive and iterated learning models, using a typology taken from an experimental study by Smith & Wonnacott (2010), and compare the results to the results observed in the experimental literature on human learning of artificial grammars.

CHAPTER 4

VARIATION BIAS: THE PLURAL MARKER TYPOLOGY

Chapter 2 demonstrated the application of the interactive and iterated learning models using a simple typological space with two constraints, which allowed for fairly straightforward analysis of the models' effects on predicting gradient typological distributions. The results of both the interactive learning model simulations (described in §2.2) and the iterated learning model simulations (described in §2.3) showed a bias against gang effect patterns and a bias away from variation in the grammar. In Chapter 3, I further investigated the models' bias against gang effect patterns, using the typology of possible contrast patterns between [s] and [ʃ]. The results of both of the learning models greatly reduced the predicted probability of the gang effect Elsewhere Neutralization pattern, which mirrors the observed rarity of this pattern type across natural languages.

In this chapter, I now turn to the bias against variability in the grammar, and disambiguate between two types of variable behavior in the grammar: *variation* in the sense it has been used throughout the dissertation so far, where multiple output realizations are possible for a given input, and *exceptionality*, where phonologically similar morphemes display arbitrarily different behavior. I investigate the predictions of the interactive and iterated learning models for patterns of variation and exceptionality by modeling a typology based on artificial language learning experiments, in which human participants show a bias away from variation, towards more deterministic patterns, but not necessarily away from patterns of exceptionality, towards more consistent patterns. While the results of the iterated and interactive learning models show a bias away from variation, mirroring the observed bias in humans, they also show a strong bias away from patterns of exceptionality, leaving open questions relating the strength of this bias in the models' behavior to human behavior and empirical typology.

4.1 Introduction

There has been much previous work on examining the cross-linguistic typology of different pattern types, and several large-scale databases have been created to compile this information in a searchable, centralized form, for example P-base (Mielke, 2008) and the World Atlas of Linguistic Structures (WALS; Dryer & Haspelmath, 2013). However, despite an increasing number of studies aimed at accounting for variation in phonology (see Coetzee & Pater, 2011, for an overview), and a rich sociolinguistic literature on patterns of variation and exceptionality (see e.g. Kelly, 1988; Sonderegger & Niyogi, 2013; Labov, 1994, 2001, 2010), it is not yet feasible to construct a reliable estimate of the relative frequencies of deterministic patterns and patterns of variation cross-linguistically. Collecting, categorizing, and compiling the available data into a database from which it would be possible to produce such a typological estimate is a task outside of the scope of the current work. For empirical evidence about biases in humans, then, I will instead rely on the body of experimental literature which investigates how humans in a laboratory setting learn from artificial grammars exhibiting variation.

Before reviewing this literature, it is first important to disambiguate between different senses in which the term "variation" has been used across different works, and to establish the terminology that will be used throughout the rest of this chapter. The term *variation* itself has been used throughout the dissertation so far to refer to cases in which there are multiple possible output forms for a given input, and each of these outputs has a different associated probability. Different works have added qualifiers such as “free”, “unpredictable”, or “unconditioned” variation to refer specifically to cases where the likelihood of a particular output remains consistent across contexts, in order to distinguish from cases where the likelihood of a particular output does vary across contexts, referred to as “predictable” or “conditioned” variation (though it has been argued that variation is usually, if not always, conditioned in some way, see e.g. Coetzee & Pater, 2011). One example of conditioned variation is *t/d*-deletion in American English (see e.g. Labov, 1989; Guy, 1994; Coetzee, 2004), where deletion of a word-final [t] or [d] occurs probabilistically, but is more likely in unstressed syllables and three-consonant clusters, and less likely before vowels or with the past tense suffix (e.g. in *missed*).

Another type of variable behavior is *exceptionality*, which refers to cases where different morphemes or lexical items arbitrarily display different behavior, despite being phonologically similar. This is opposed to consistent patterns, where phonologically similar forms display the same behaviors. The variable behavior in cases of exceptionality is observed across the lexicon, then, rather than within individual items. One example of this comes from English stress, where three-syllable words with a final schwa can have penultimate stress (e.g. *banána*) or antepenultimate stress (e.g. *Cánada*), and where each individual word is pronounced in a single way that cannot be predicted from its phonological properties (Moore-Cantwell, 2015). Cases of exceptionality can also display variation, where the exceptional behavior manifests as different probabilistic trends, rather than different deterministic choices.

4.1.1 Experimental background

While many artificial language learning studies focus on comparing participants’ ability to learn patterns with different formal structures or phonetic substance (see Moreton & Pater, 2012a,b, for an overview), there are a number of studies which investigate how people, both children and adults, learn from training data which contains patterns of variation and patterns of exceptionality (Reali & Griffiths, 2009; Hudson Kam & Newport, 2005, 2009; Wonnacott, 2011; Hudson Kam, 2015; Samara et al., 2017, among others).

Some of these studies have tested both children and adults on learning patterns of free variation in a laboratory setting (Hudson Kam & Newport, 2005, 2009; Hudson Kam, 2015; Wonnacott, 2011; Culbertson et al., 2012). In Hudson Kam & Newport (2005), participants were trained on an artificial language in which nouns probabilistically occurred with or without a determiner in a pattern of free variation. Participants listened to sentences in this language, which accompanied and described visual scenes, and then were tested on their ability to produce sentences in response to a visual scene. The results of this study found that adults tended to display probability-matching behavior, where the frequencies at which they used determiners in their own productions matched the frequencies observed in their training data, while children tended to display regularizing behavior, where the pattern of determiner use (or omission) in their productions was more deterministic than observed in their training data. While children are more likely to regularize, other studies have shown that adults will regularize free variation under some circumstances. Wonnacott & Newport (2005) showed that adults will regularize variation in determiner usage and word order when tested on novel vocabulary items that were not found in their training data, and Hudson Kam & Newport (2009) showed that adults were more likely to regularize variation in determiner usage when there were more competing alternatives in free variation. In addition to free variation, it has been shown that participants can learn conditioned variation in an experimental setting. Samara et al. (2017) trained both adults and children on an artificial language in which nouns were obligatorily followed by one of two meaningless particles, and where particle use was conditioned either deterministically or probabilistically on speaker identity. In both conditions, both adults and children were able to acquire the speaker identity cue conditioning determiner use.

Other studies have tested both children and adults on learning patterns of exceptionality in a laboratory setting (Wonnacott et al., 2008; Wonnacott, 2011; Hudson Kam, 2015; Samara et al., 2017). Wonnacott et al. (2008) trained adult participants on an artificial language in which verbs could occur in one of two word orders: Verb Subject Object, or Verb Object Subject Particle. In one experiment (Experiment 3 in their paper), they compared the generalization behavior between participants in two conditions. In the "generalist" condition, participants were trained on a pattern of free variation in which any verb could be used with either word order, but verbs occurred in the Verb Object Subject Particle order seven times more often. In the "lexicalist" condition, participants were trained on a pattern of exceptionality in which one verb only occurred in the Verb Subject Object order, and the other seven verbs only occurred in the Verb Object Subject Particle order. After training, participants were given a production test, in which they were tested using novel verbs, and a grammaticality judgment test. In both conditions, participants were more likely to use a novel verb in the Verb Object Subject Particle order, which was the most frequent construction observed in the training data, showing that participants in both conditions are sensitive to overall lexical statistics. In the grammaticality judgment test, participants in the lexicalist condition gave higher grammaticality ratings to trained verbs used in the same word order as they were used in the training data, while participants in the generalist condition accepted trained verbs used in either word order, showing that participants are sensitive to the difference between exceptionality and variation, even though they displayed similar behavior in generalizing to novel verbs. In a follow-up study, Wonnacott (2011) showed that children display a similar pattern of results as the adults in Wonnacott et al. (2008) using a similar experimental paradigm.

In sum, the experimental evidence suggests that patterns of both variation and exceptionality are learnable by both children and adults, and that while both age groups show a bias towards regularizing variation, this bias is stronger in children than in adults. This bias against variation is in line with the bias towards more deterministic grammars observed in earlier chapters, however, comparing these experimental results to the results of the interactive and iterated learning models is not entirely straightforward. The structure of the majority of the experimental studies listed here evaluates learning by an individual from some target grammar, analogous to one generation of the iterated learning model. It is not clear how one would extrapolate from these experimental results to predictions about the effect of interaction between individuals, as in the interactive learning model, or to predictions about the cumulative effect of these biases when transmitting patterns across multiple individuals in a chain, as in the iterated learning model. There are, however, a small number of artificial grammar learning studies which use an iterated learning experimental paradigm, and one example which uses this paradigm to test participants on learning variation and exceptionality (Smith & Wonnacott, 2010), which allows for a more straightforward comparison between experimental results and modeling results, and so will be used as the basis of comparison for the investigation of the predictions of the learning models.

4.1.1.1 Smith & Wonnacott (2010)

In Smith & Wonnacott (2010), the authors conduct an artificial language learning experiment with the aim of investigating whether predictability in natural language might be a consequence of constraints inherent in language acquisition, amplified through cultural transmission of a pattern across generations of learners. They tested this hypothesis by organizing their participants into short iterated learning chains in which the output generated by one participant was used as the input for the next participant in the chain. They then analyzed whether and how the pattern changed as it was transmitted across participants. They had fifty participants, who were organized into 10 learning chains of five participants.

In the experiment, participants were tasked with learning how to form the plurals of nouns in an artificial grammar. The grammar consisted of four English nouns (cow, pig, rabbit, and giraffe), which could be marked as plural using one of two plural markers (fip and tay), and a verb glim which meant "to move". Participants were trained on this artificial grammar by observing sentences in text, paired with visual scenes depicting one or a pair of cartoon animals performing a "move" action, and then retyping the text description of the scene. For example, a scene depicting two cows moving could be described as "glim cow tay". There were eight possible scenes, a singular and a plural scene for each of the four nouns, which they saw once in each of twelve training blocks. Participants were then tested by observing each scene, and typing the description. They saw each of the eight scenes once in each of four testing blocks.

The first participant in each chain was trained on a pattern in which each marker could be used with each noun, but one marker was used three times as often. Five chain-initial participants were trained on a pattern where, for each word, the plural was marked with fip 75% of the time, and the other five were trained on a pattern where, for each word, the plural was marked with tay 75% of the time. Thus, while one marker is more common, the variation between the use of each of the markers displays the same probabilistic tendency on each noun. Each subsequent participant in a learning chain was trained on the responses given by the previous participant during testing; the previous participant's responses in their four testing blocks were repeated three times to yield twelve training blocks for the next participant.

In the results of their experiment, Smith & Wonnacott (2010) find that, as the pattern is transmitted across the learning chains, participants gradually regularize the variation observed in their input. The patterns in each chain were made more deterministic either by eliminating one of the plural markers, or by indexing the use of each plural marker to a subset of lexical items in a pattern of exceptionality. Figure 4.1 reproduces the figures depicting the experimental results from Smith & Wonnacott (2010). The left-hand figure shows, for each participant in each chain, the number of plurals marked in testing using the initial majority marker, ranging from a minimum of zero to a maximum of sixteen. Each dashed line corresponds to a different chain, while the solid line shows the mean. In this figure it can be seen that the behavior of the learning chains diverges over time, with some chains converging on consistent plural marking across nouns, using the initial majority marker for zero or sixteen of the scenes in the testing blocks, and other chains producing a range of behavior in between, using the initial majority marker around 25%, 50%, or 75% of the time.

The participants are not producing patterns of variation, however, as can be seen in the right-hand figure in Figure 4.1. This graph shows, for each position in the learning chain, the mean conditional entropy of the patterns produced by each participant, averaged over all ten learning chains. Conditional entropy is a measure indicating how deterministic a pattern is, as discussed in §1.3.3. A conditional entropy value of zero corresponds to a completely deterministic pattern, while greater and greater values indicate systems which are less and less predictable. As can be seen in the graph, the conditional entropy decreases as the patterns are transmitted across the learning chains. This shows that, even though some of the chains preserve the use of both plural markers, these chains have developed patterns of deterministic exceptionality, rather than preserving the variation from the original pattern in the chain.

The experimental results from Smith & Wonnacott (2010) thus provide evidence for a bias against variation in human learners. The participants in the experiment used different strategies to regularize variation in plural marking when learning the artificial language, including eliminating one of the plural markers, or developing

Figure 4.1. Figures showing the experimental results from Smith & Wonnacott (2010), reproduced from that paper

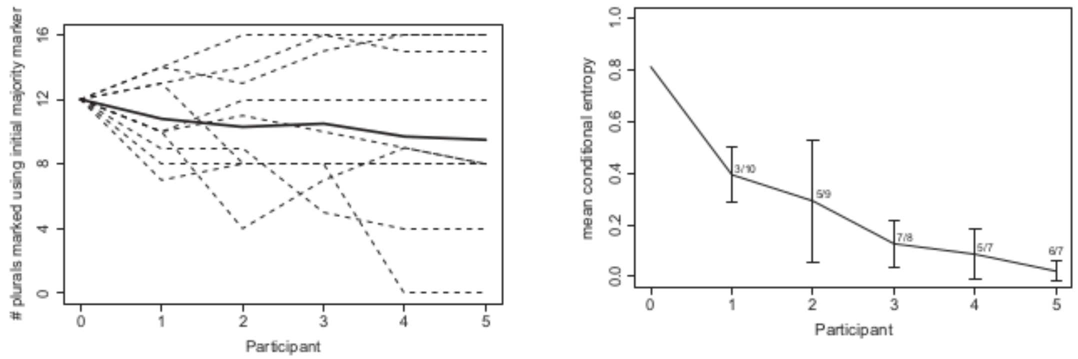


Fig. 1. Number of plurals marked using the marker which was initially in the majority in each chain (out of 16 two-animal scenes encountered by each participant during testing). Solid line gives mean, dashed lines show individual chains. Participant 0 is the experimenter-designed input language used to train the first participant in each chain.

Fig. 2. Conditional entropy of the language produced by each participant (participant 0 is the input language), averaged over all 10 chains. Error bars give 95% confidence intervals on the mean. Annotations give the number of languages which have significantly non-random use of marking (see text for details) as a proportion of those languages which still use multiple markers for the plural.

patterns of exceptionality, where each plural marker is indexed to a subset of lexical items. In previous chapters, I have shown that both the iterated and interactive learning models display an emergent bias away from variability in the grammar, tending towards accumulating probability on one output candidate over its competitors. The typologies used in previous chapters, however, have not distinguished between variation, where a particular input can have multiple different output realizations, and exceptionality, where different words arbitrarily display different behavior. In order to further explore the models' biases regarding variation and exceptionality, I applied them to a typology using the artificial grammar in Smith & Wonnacott (2010).

4.1.2 Variation typology

In order to model the artificial grammar from Smith & Wonnacott (2010) using a MaxEnt grammar, a constraint set is needed which is capable of representing both consistent patterns, in which one plural marker is used consistently across lexical items, and patterns of exceptionality, in which each lexical item can select a plural marker according to its own individual preference. The constraint set chosen here makes use of lexically-indexed constraints (see Pater, 2000; Prince & Smolensky, 1993/2004), which are specific versions of general constraints, whose domain of application is restricted to a particular morpheme. Lexically-indexed constraints have been used in previous work to account for a number of phenomena, including lexically-conditioned stress patterns (Pater, 2000), morphemes that exceptionally block or trigger a phonological process (Pater, 2010), and gradient productivity in processes with exceptions Moore-Cantwell & Pater (2016).

The MaxEnt grammar used to model the Smith & Wonnacott (2010) experiment contains four possible input lexical items, labeled "cow", "pig", "rabbit", and "giraffe", and two possible ways of marking the plural for each word, "fip" and "tay". Thus, for example, the two possible outputs for "cow" are "cow fip" and "cow tay". There are 10 constraints in this system, listed in Table (4.1): two general constraints, one preferring "fip" as the plural marker in all cases and one preferring "tay", as well as lexically-indexed versions of these constraints which only apply to a specific lexical item (2 markers × 4 words = 8 lexically-indexed constraints). The lexically-indexed constraints allow the grammar to represent patterns in which different words condition the use of different plural markers. While using the lexically-indexed constraints alone would allow the grammar to represent patterns of consistent plural marker use across lexical items, as a case where all of the lexical items happen to prefer the same marker, the general constraints are useful in representing properties of the overall lexical statistics, as well as for generating predictions for forms not indexed by a lexically-indexed constraint, or for novel forms not already in the lexicon. Additionally, much prior work using lexically-conditioned constraints assumes that they exist in addition to general constraints, rather than in isolation (Pater, 2000, 2010, 2012; Moore-Cantwell & Pater, 2016). The general constraints are also crucial in generating a bias towards consistent plural marking across lexical items, as discussed further in §4.4.1

Table 4.1. Constraints used to model the plural marker typology

Pl=fip: Assign one violation if the plural is not marked with the plural marker “fip”

Pl=tay: Assign one violation if the plural is not marked with the plural marker “tay”

Pl(W_x)=fip: Assign one violation if the plural of W_x is not marked with the plural marker “fip”

Pl(W_x)=tay: Assign one violation if the plural of W_x is not marked with the plural marker “tay”

This grammatical space is illustrated in Tableau 4.1, showing the complete list of constraints, the four input lexical items, their corresponding output forms, and their constraint violations. Each output form violates one of the general constraints as well as one of the corresponding lexically-specific constraints, e.g. “cow fip” violates Pl=tay and Pl(cow)=tay. The lexically-specific constraints only apply to their specified lexical item, and are otherwise vacuously satisfied.

Tableau 4.1. Tableaux showing the grammar space for the plural marker typology

	Pl=fip	Pl=tay	Pl(cow)=fip	Pl(cow)=tay	Pl(pig)=fip	Pl(pig)=tay	Pl(rabbit)=fip	Pl(rabbit)=tay	Pl(giraffe)=fip	Pl(giraffe)=tay
/cow/	–	–	–	–	–	–	–	–	–	–
cow fip		-1		-1						
cow tay	-1		-1							
/pig/	–	–	–	–	–	–	–	–	–	–
pig fip		-1				-1				
pig tay	-1				-1					
/rabbit/	–	–	–	–	–	–	–	–	–	–
rabbit fip		-1						-1		
rabbit tay	-1						-1			
/giraffe/	–	–	–	–	–	–	–	–	–	–
giraffe fip		-1								-1
giraffe tay	-1								-1	

Given these constraints, inputs, and possible corresponding outputs, there are sixteen possible categorical patterns, listed in Table 4.2. These sixteen patterns represent all possible combinations of plural marker preferences on each noun. For the purposes of the investigation in this chapter, however, the matter of interest is whether or not the grammar displays lexically-conditioned use of plural markers. Thus, it is more useful to abstract away from the identity of the particular lexical items preferred and instead categorize each pattern according to how many lexical items prefer to take one plural marker, and how many prefer the alternative. This yields three pattern types, one in which each lexical item prefers the same plural marker (4X0Y), one in which three lexical items prefer one marker and the remaining lexical item prefers the other (3X1Y), and one in which half of the lexical items prefer one marker and the other half prefer the other (2X2Y).

Table 4.2. The set of possible patterns in the plural marker typology

	Pattern Type	Marked with FIP	Marked with TAY
1	4X0Y	COW, PIG, RABBIT, GIRAFFE	∅
2	3X1Y	PIG, RABBIT, GIRAFFE	COW
3	3X1Y	COW, RABBIT, GIRAFFE	PIG
4	3X1Y	COW, PIG, GIRAFFE	RABBIT
5	3X1Y	COW, PIG, RABBIT	GIRAFFE
6	2X2Y	RABBIT, GIRAFFE	COW, PIG
7	2X2Y	PIG, GIRAFFE	COW, RABBIT
8	2X2Y	PIG, RABBIT	COW, GIRAFFE
9	2X2Y	COW, GIRAFFE	PIG, RABBIT
10	2X2Y	COW, RABBIT	PIG, GIRAFFE
11	2X2Y	COW, PIG	RABBIT, GIRAFFE
12	3X1Y	COW	PIG, RABBIT, GIRAFFE
13	3X1Y	PIG	COW, RABBIT, GIRAFFE
14	3X1Y	RABBIT	COW, PIG, GIRAFFE
15	3X1Y	GIRAFFE	COW, PIG, RABBIT
16	4X0Y	∅	COW, PIG, RABBIT, GIRAFFE

Before examining the effects of applying the interactive and iterated learning models to this typology, it is necessary to first establish a baseline distribution without learning. In order to estimate the baseline distribution, 10,000 sets of constraint weights were sampled from a uniform distribution with the range 0.0-10.0, and the resulting grammars were categorized into the pattern types listed in Table 4.2, and the conditional entropy of each grammar was binned into ten ranges of values, with 0.0-0.1 as the lowest, most deterministic bin, and 0.9-1.0 being the highest, most variable bin. The baseline distribution over pattern types shows a dispreference for 2x2y patterns, with only 20% of sampled sets of weights corresponding to this pattern type, compared to 40% each for 3x1y and 4x0y patterns. The baseline distribution over conditional entropy values shows a tendency towards the lower range of values, and thus towards more deterministic grammars, with 84% of sampled grammars having a conditional entropy value lower than 0.5. The spread of the distribution is fairly even across this lower range of values, however, and doesn't show a strong pull towards the most deterministic grammars.

Table 4.3. The sampled baseline distribution over pattern types and conditional entropy values for the plural marker typology

	Pattern Type		
	4x0y	3x1y	2x2y
Sampled Baseline	0.40	0.40	0.20

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00

It is worth noting that collapsing the individual patterns into pattern type categories obscures the probability of each individual pattern, as each category contains a different number of individual patterns. The 2x2y pattern type category, for example, contains 6 individual patterns, yielding a probability of about 3% on each, while the 3x1y pattern type category contains 8 individual patterns, yielding a probability of about 5% on each. This additionally means that, while the 4x0y and 3x1y pattern type categories have equal probability in the baseline distribution, the probability of each individual 4x0y pattern is greater than that of individual patterns in other categories, at 20% ($0.40 \div 2 = 0.20$). For the purposes of the investigation in this chapter, however, there is no significant information gained by reporting the probabilities of each individual pattern, and so the analysis will instead focus on the probabilities of the pattern type categories.

The next sections present the results of applying the interactive and iterated learning models to the plural marker typology, based on Smith & Wonnacott (2010). §4.2 provides a general investigation of the predictions of the interactive learning model and the iterated learning model, while §4.3 models the experimental conditions in Smith & Wonnacott (2010) using the iterated learning model. The results of both models and both modeling scenarios show a robust bias towards consistent plural marking across lexical items, away from exceptionality, as well as a bias away from variability that is less robust across modeling conditions. As discussed in §4.3, these results contrast with the experimental findings in Smith & Wonnacott (2010), where participants showed a strong bias away from variability, and a weaker bias away from patterns of exceptionality. §4.4 discusses the effects of some parameter manipulations on the modeling results, while §4.5 provides an overall discussion and concludes the chapter.

4.2 Modeling the general case

Before modeling the specific experimental conditions implemented in Smith & Wonnacott (2010), it is important to first explore the behavior of the interactive and iterated learning models on the plural marker typology in the general case. This will establish a context for interpreting the results of modeling the experiment conditions, so that it is possible to distinguish between results that reflect general properties of the models' behavior, and results that are dependent on the particularities of the experiment conditions. I first detail the results from the interactive learning model, in §4.2.1. In §4.2.2, I detail the results from the iterated learning model, where the initial generation of learner agents were trained on grammars generated by randomly sampling constraint weights from a uniform distribution between 0-10. Both models show a bias towards consistent plural marking across lexical items, away from patterns of exceptionality, as well as a bias away from variability in the grammar, though this bias is less consistent across modeling conditions.

For both models, I test two weight initialization conditions for learner agents: one in which constraints are initialized with random weights sampled from a uniform distribution with the range 0-10 (the random-weight initialization condition), giving learners a starting point with a random set of preferred output forms, and one in which constraints are initialized with a weight of five (the five-weight initialization condition), giving learners a starting point where they have an equal probability on each competing output form. The five-weight initialization condition in this chapter serves the same purpose as the zero-weight initialization condition used in Chapter 2 and Chapter 3. Because each pair of competing output candidates in the plural marker typology has a complementary set of constraint violations (see Tableau 4.1), this condition yields equal probability on all output candidates, as was the case for the zero-weight initialization condition. The symmetric nature of the constraint set used for this typology, however, results in interference during learning when the constraint weights are initialized at zero and there is a lower bound on constraint weights at zero, as is assumed throughout this dissertation. To avoid any issues that may arise from this interference, then, the zero-weight condition used in previous chapters is replaced in this chapter with the five-weight condition, where five is an arbitrarily chosen positive real number. See §4.4.3 for a more detailed discussion of this issue.

4.2.1 Interactive learning simulations

In this section, I present the results of applying the interactive learning model to the plural marker typology based on the artificial grammar used in Smith & Wonnacott (2010), described in §4.1. In all simulations, both learning agents were initialized with the same set of constraint weights, with the starting weights determined by the initialization condition. For simulations in both initialization conditions, 10,000 runs were performed, where the learning agents exchanged data for 10,000 learning steps, with a learning rate of 0.1.

Table 4.4 summarizes the distribution over pattern types and conditional entropy values across all runs of the simulations. The table shows the results for both initialization conditions, as well as repeating the sampled baseline distribution. Both initialization conditions show a stronger bias towards consistent plural marking across lexical items when compared to the baseline distribution. This bias is stronger in the five-weight condition, where 4X0Y patterns emerge in 79% of runs, than in the random-weight condition, where 4X0Y patterns emerge in 54% of runs (compare to 40% of sampled sets of weights in the baseline distribution). Both initialization conditions additionally show a bias towards more deterministic patterns, where agents in about 65% of runs have learned grammars with conditional entropy values less than 0.1, compared to only 18% of sampled sets of weights in the baseline distribution.

The distribution over pattern types in both conditions is examined in more detail in Figure 4.2, which shows the change in the proportion of runs corresponding to each pattern type across learning steps in the simulation. The five-weight condition is plotted in the graph on the left, while the random-weight condition is plotted in the graph on the right. Both initialization conditions immediately show a preference for the 4X0Y patterns as the simulation begins. In the five-weight condition, however, this preference is marked by a large, sharp increase in the proportion of runs corresponding

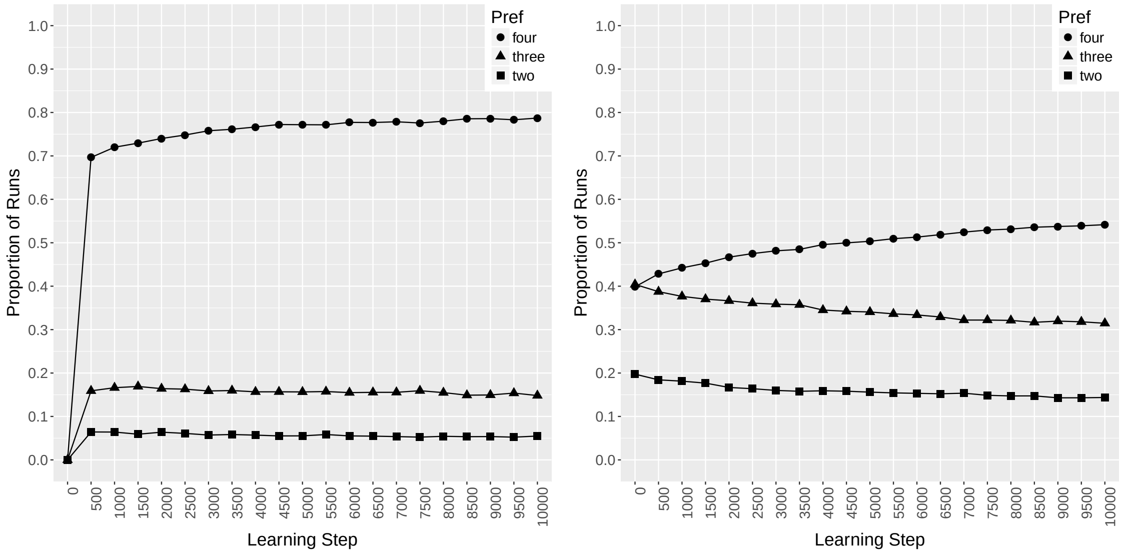
Table 4.4. The distribution over pattern types and conditional entropy after learning with the interactive learning model in the plural marker typology, in the general case

	Pattern Type		
	4X0Y	3X1Y	2X2Y
Sampled Baseline	0.40	0.40	0.20
Five-wght Cond.	0.79	0.15	0.08
Rand-wght Cond.	0.54	0.32	0.14

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Five-wght Cond.	0.65	0.15	0.09	0.06	0.03	0.01	0.00	0.00	0.00	0.00
Rand-wght Cond.	0.66	0.19	0.10	0.03	0.01	0.00	0.00	0.00	0.00	0.00

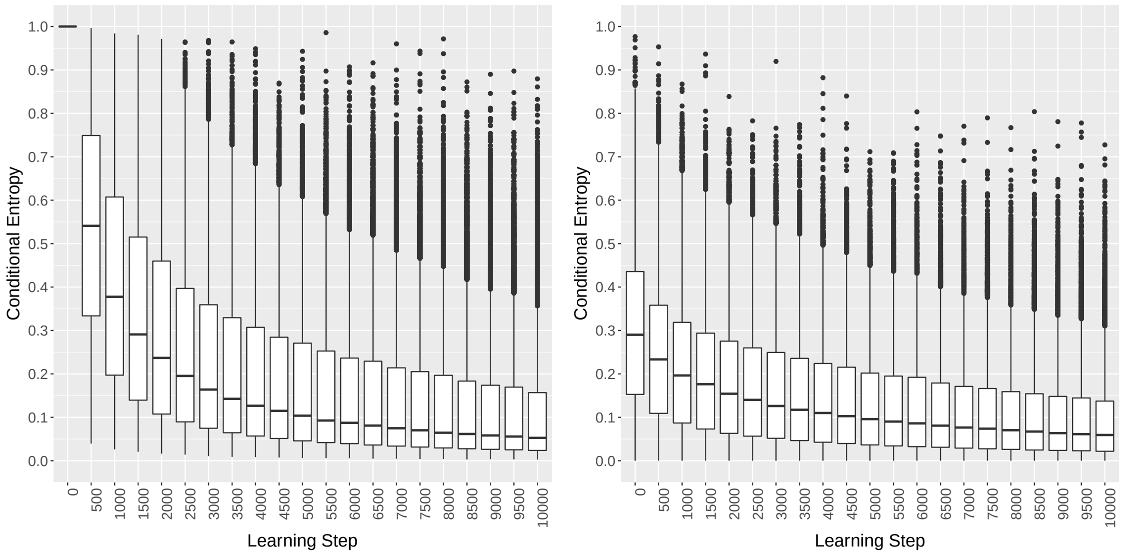
to the 4X0Y pattern, which then continues to build slowly, while the preference in the random-weight condition increases gradually across learning steps.

Figure 4.2. Plots showing the change in the distribution over pattern types across learning steps with the interactive learning model in the plural marker typology, in the general case (Left: five-weight condition, Right: random-weight condition)



The distribution over conditional entropy values in both conditions is examined in more detail in Figure 4.3, which shows the change in conditional entropy across learning steps in the simulation. The five-weight condition is plotted in the graph on the left, while the random-weight condition is plotted in the graph on the right. Both initialization conditions show a steady drift towards more deterministic grammars, which can be observed in both the decreasing mode and variance of the distributions over conditional entropy values across learning steps.

Figure 4.3. Plots showing the change in conditional entropy across learning steps with the interactive learning model in the plural marker typology, in the general case (Left: Five-weight condition, Right: Random-weight condition)



4.2.2 Iterated learning simulations

In this section, I present the results of applying the iterated learning model to the plural marker typology, examining the model's behavior both within one generation of agents, and across multiple generations of agents. In all simulations, the initial generation of teacher agents have grammars initialized by randomly sampling constraint weights. The starting weights for new learner agents introduced into the simulation were determined by the initialization condition. For simulations in all conditions, 10,000 runs were performed, with a learning rate of 0.1.

4.2.2.1 Within one generation

I will first discuss the results of examining one generation of the iterated learning model, in which the randomly sampled initial target grammars are transmitted to the first generation of learner agents. In this set of simulations, the learner agents received data from their target distribution for 1,000 learning steps. Table 4.5 summarizes the distribution over pattern types and conditional entropy values across all runs of the simulations, showing the results for both initialization conditions, as well as repeating the sampled baseline distribution. In both initialization conditions, the results after 1,000 learning steps are only minimally different from the sampled baseline distribution, over both pattern types and conditional entropy values. Because the target grammars are randomly sampled, the distribution over target grammars is the same as the sampled baseline distribution. Thus, after 1,000 learning steps, the learner agents have reached a distribution that has very similar properties to the target distribution.

Table 4.5. The distribution over pattern types and conditional entropy after one generation of learning with the iterated learning model in the plural marker typology, in the general case

	Pattern Type		
	4x0Y	3x1Y	2x2Y
Sampled Baseline	0.40	0.40	0.20
Five-wght Cond.	0.39	0.38	0.19
Rand-wght Cond.	0.40	0.40	0.20

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Five-wght Cond.	0.14	0.14	0.19	0.21	0.15	0.10	0.04	0.02	0.00	0.00
Rand-wght Cond.	0.16	0.14	0.20	0.20	0.14	0.10	0.04	0.02	0.00	0.00

Figure 4.4 plots the change in the proportion of runs corresponding to each pattern type across learning steps in the simulations. The five-weight condition is shown on the left, while the random-weight condition is shown on the right. In both initial-

ization conditions, the distribution seems to reach a fairly stable state by 400 learning steps, with agents in the five-weight condition possibly reaching a stable distribution sooner, around 250 learning steps. In the five-weight condition, there is an initial increase in the proportion of runs corresponding to a 4X0Y pattern, which then decreases as the proportion of agents who have learned 3X1Y or 2X2Y patterns increases. This initial excess of 4X0Y patterns is due to cases where agents learning a 3X1Y or 2X2Y target pattern have not yet increased the weights of the lexically-indexed constraints high enough relative to the general constraints to overcome the pressure for consistent use of one plural marker across all lexical items. In the random-weight condition, on the other hand, there is an initial decrease in the proportion of runs corresponding to 4X0Y patterns, and an increase in 3X1Y and 2X2Y patterns, which then reverses – the proportion of 4X0Y patterns increases while the proportion of 3X1Y and 2X2Y patterns decreases – before the distribution settles into an approximation of the target distribution. The differences between initialization conditions are due to differences in potential directions for mislearning, as discussed in §4.2.3.

Figure 4.4. Plots showing the change in the distribution over pattern types across one generation of the iterated learning model in the plural marker typology, in the general case (Left: five-weight condition, Right: random-weight condition)

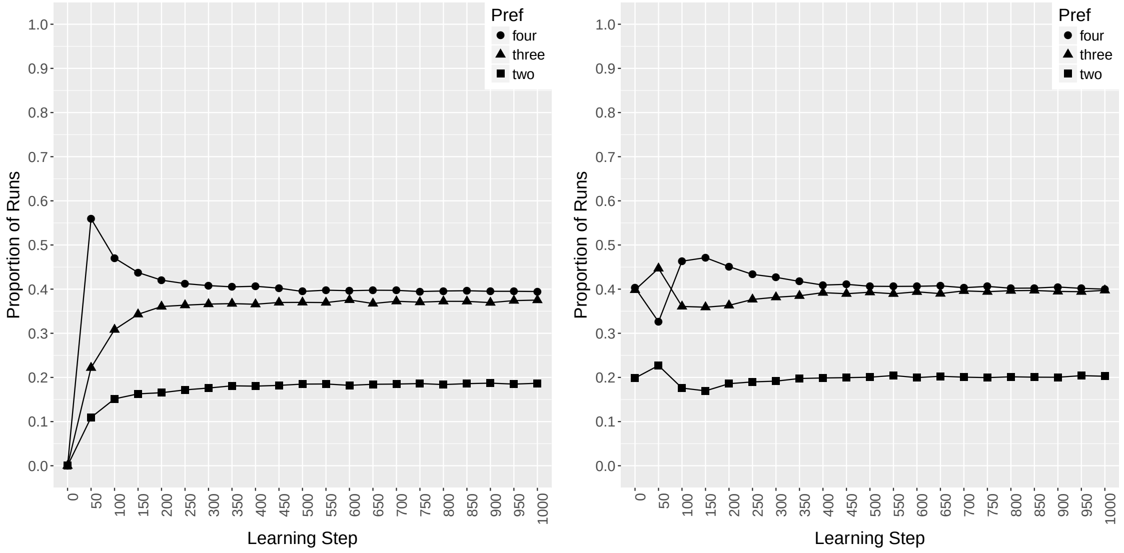
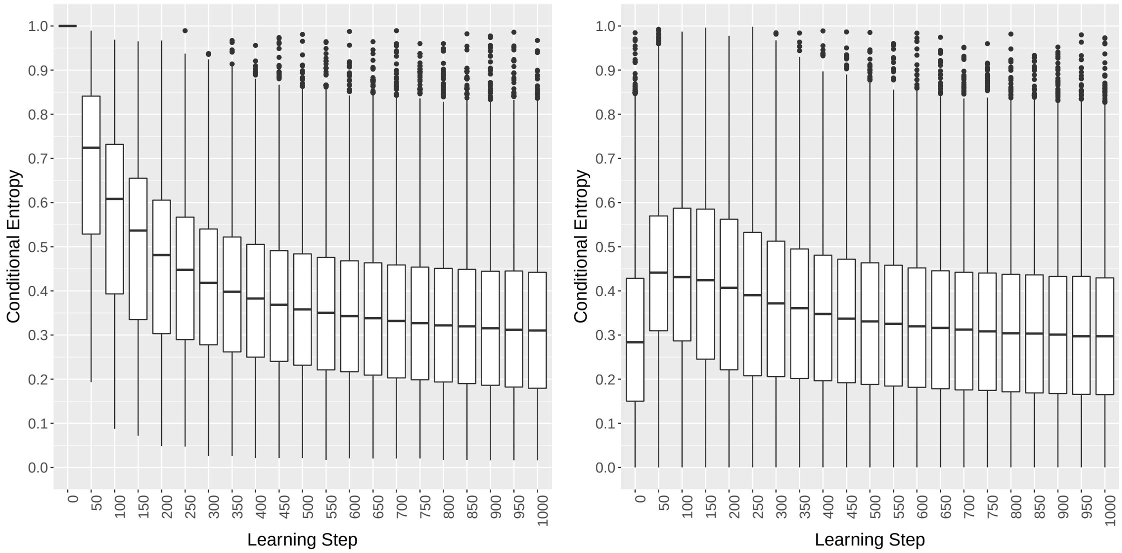


Figure 4.5 plots the change in the distribution over conditional entropy values across learning steps in the simulations. In both initialization conditions, the distributions reach a state very similar to the target distribution, which has similar properties to the sampled baseline distribution. In the five-weight condition, the mode of the conditional entropy distribution falls from its initial value of 1, which corresponds to the initial state of the learners, where every output candidate has equal probability. In the random-weight condition, the mode of the conditional entropy distribution first rises, and then falls to fit the target distribution. This initial rise is due to the fact that learner agents who are fitting a target pattern different from the one corresponding to their initial state must move through states of higher variation as they adjust their constraint weights to fit their target.

Figure 4.5. Plots showing the change in conditional entropy across one generation of the iterated learning model in the plural marker typology, in the general case (Left: five-weight condition, Right: random-weight condition)



4.2.2.2 Across generations

In order to observe any learning biases emergent in the iterated learning model across multiple generations, it is necessary to truncate the number of learning steps allowed to each agent at a point before the learning agents are able to fully fit their target grammars. For the simulations presented in this section, the learning period for each agent was 400 learning steps, and each run iterated over 50 generations.

Table 4.6 summarizes the resulting distributions over pattern types and conditional entropy values, showing the results from both the five-weight and random-weight conditions, as well as repeating the sampled baseline distribution. Both initialization conditions show a bias for consistent plural marking across lexical items, with over 75% of runs ending in a 4x0y pattern, compared to 40% of sampled sets of weights in the baseline distribution. However, only the random-weight condition shows a clear bias towards more deterministic patterns, with 27% of runs ending in grammars with a conditional entropy value less than 0.1, and 32% of runs ending in grammars with a conditional entropy value between 0.1-0.2, compared to 18% and 14% respectively in the baseline distribution. The five-weight condition on the other hand, has more runs ending in grammars with a conditional entropy value between 0.1-0.2 (27%), but fewer ending in values less than 0.1 (6%) and comparatively more ending in values greater than 0.6.

Table 4.6. The distribution over pattern types and conditional entropy after 50 generations of learning with the iterated learning model in the plural marker typology, in the general case

	Pattern Type		
	4x0y	3x1y	2x2y
Sampled Baseline	0.40	0.40	0.20
Five-wght Cond.	0.76	0.14	0.06
Rand-wght Cond.	0.78	0.15	0.06

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Five-wght Cond.	0.06	0.27	0.16	0.10	0.10	0.09	0.11	0.07	0.03	0.01
Rand-wght Cond.	0.27	0.32	0.11	0.09	0.08	0.07	0.04	0.02	0.00	0.00

Figure 4.6 shows the change in the distribution over pattern types in more detail, plotting the proportion of runs corresponding to each pattern type across generations

in the simulations. The five-weight condition is shown on the left, and the random-weight condition is shown on the right. Both initialization conditions show a very similar trajectory across generations, with the proportion of runs in 4X0Y patterns rising, and the proportions of runs in other pattern types falling across generations.

Figure 4.6. Plots showing the change in the distribution over pattern types across 50 generations of the iterated learning model in the plural marker typology, in the general case (Left: five-weight condition, Right: random-weight condition)

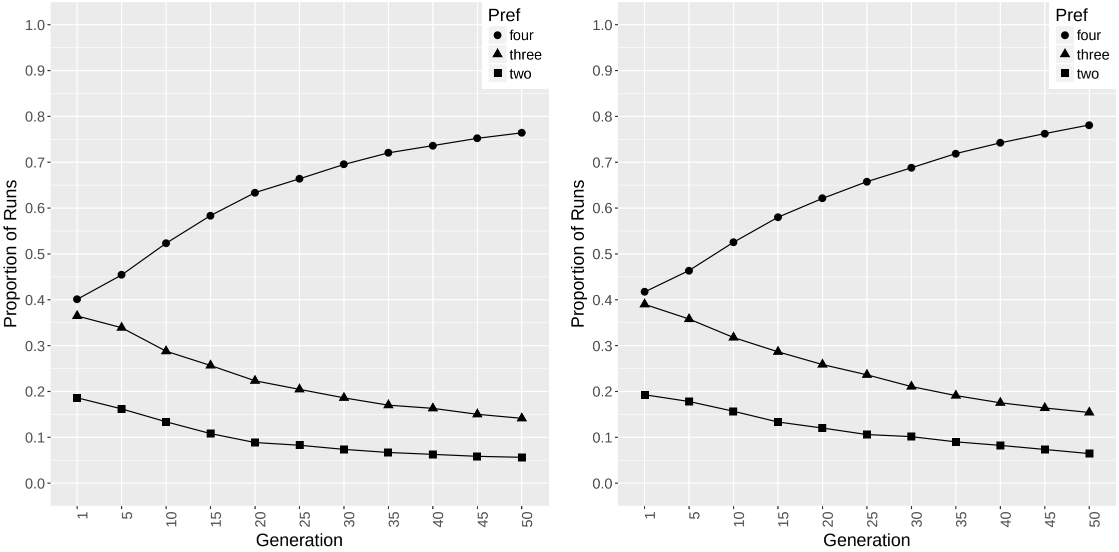
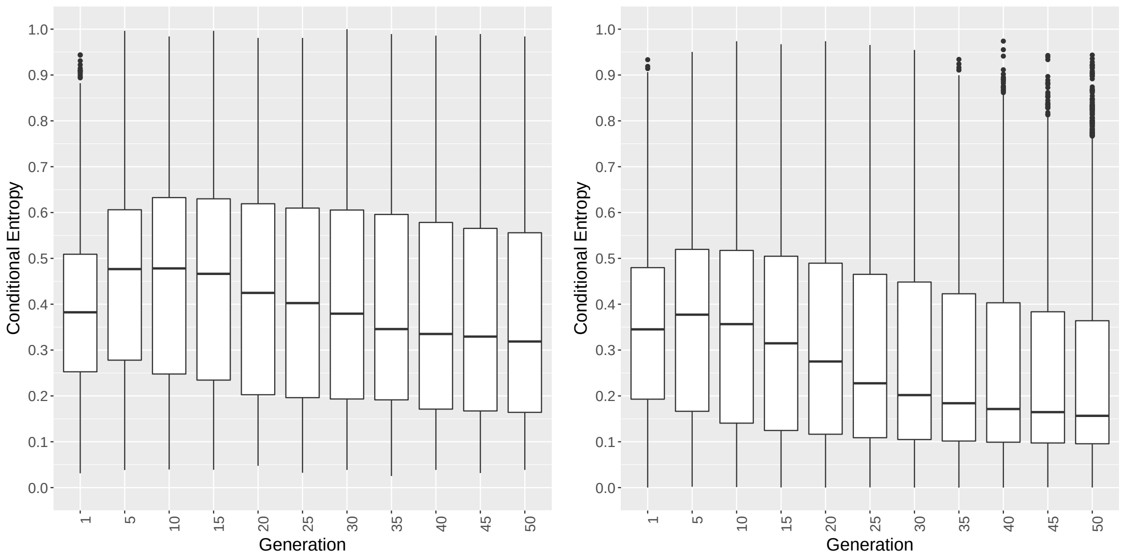


Figure 4.7 shows the change in the distribution over conditional entropy values in more detail, plotting the distributions over conditional entropy values across generations of the simulations. Both initialization conditions show an initial increase in the mode and variance of the distribution, around 5-10 generations, which then both decrease as the simulation continues. However, the decrease in mode and variance is more significant in the random-weight condition than for the five-weight condition.

Figure 4.7. Plots showing the change in conditional entropy across 50 generations of the iterated learning model in the plural marker typology, in the general case (Left: five-weight condition, Right: random-weight condition)



4.2.3 General case discussion

§4.2 presented an investigation of the predictions of the interactive and iterated learning models for the plural marker typology in the general case. For the interactive learning model, I tested two initialization conditions: one in which agents were initialized with all constraint weights set to the arbitrary positive real number five (the five-weight condition), and one where all constraint weights were randomly sampled from a uniform distribution between 0.0-10.0 (the random-weight condition). For the iterated learning model, the initial set of target grammars were generated by randomly sampled constraint weights, and simulations were performed testing the effects of the five-weight and random-weight initialization conditions for new learner agents introduced into the generational chain.

Both learning models showed a consistent bias towards preferring the same plural marker across all lexical items when compared to the sampled baseline distribution, with higher proportions of runs ending in a 4X0Y pattern in all conditions. Both models additionally showed a bias towards more deterministic grammars. However, this bias was much stronger in the interactive learning model than in the iterated learning model. In the iterated learning model, only the random-weight condition showed a clear bias towards more deterministic grammars, with higher proportions of runs ending with conditional entropy values closer to zero compared to the sampled baseline distribution. The five-weight condition showed mixed results, with a higher proportion of runs ending with conditional entropy values between 0.1-0.2, but a lower proportion ending with lower values, and a higher proportion ending in higher values, compared to the baseline distribution.

The strong bias towards consistent plural marking across lexical items is due to the presence of the general constraints preferring FIP or TAY. Each time the learning agents update their grammars, the learners promote the general constraints favoring the plural marker observed on the teacher's form, increasing the probability of that plural marker across lexical items, as well as the specific constraints favoring the particular combination of lexical item and plural marker. If the general constraints are not included in the constraint set, the bias towards consistent plural marking across lexical items does not emerge, as discussed further in §4.4.1.

The strong bias toward more deterministic grammars in the interactive learning model is due in part to the learning mechanic of mutual interaction, and in part to the error-driven learning algorithm. As the agents' grammars become more deterministic, they become less likely to sample different outputs for the same input, and thus less likely to trigger a grammar update. While it is still possible for the agents to disagree and move away from deterministic grammar states, the agents become more likely to remain at the same grammar state across multiple learning steps as their grammars become more deterministic.

The bias towards more deterministic grammars was weaker in the iterated learning model, in part because of the influence of having a target grammar, and in part because of the influence of the initialization conditions on new learner agents. Where agents in the interactive learning model are free to drift around the learning space, the agents in the iterated learning model must attempt to fit a particular grammar, no matter what pattern or how variable. This preserves less deterministic grammars for longer than would be seen in the interactive learning model. Additionally, the initialization conditions on new learners influence the directions in which learner agents mislearn their target grammars, as displayed in Table 4.7, contributing to the different results observed between the five-weight and random-weight initialization conditions.

Table 4.7 shows the transition probabilities between a target conditional entropy value in a given range, and whether the learned conditional entropy value was in a lower range, in the same range, or in a greater range. The five-weight condition is shown on the left, and the random-weight condition is shown on the right. Target grammars with conditional entropy values above 0.1 were more likely to be learned faithfully in the five-weight condition than in the random-weight condition, while target grammars with conditional entropy values below 0.1 were more likely to be learned faithfully in the random-weight condition than in the five-weight condition. This difference contributes to the slower decline in conditional entropy over generations in the five-weight condition, compared to the random-weight condition.

Additionally, when the learned grammar does not have a conditional entropy value in the same range as the target value, the direction of mislearning is more likely to be towards higher conditional entropy values in the five-weight condition, compared to the random-weight condition. This can be seen by comparing the values in the "Lower" and "Greater" columns in both conditions. Disregarding the highest and lowest ranges where only one direction of mislearning is possible, mislearning towards greater conditional entropy values is more likely for 5/8 target ranges in the five-weight condition, compared to only 3/8 target ranges in the random-weight condition. This difference contributes to the slower decline in conditional entropy values across generations in the five-weight condition, as well as the greater proportion of grammars with higher conditional entropy values, compared to the random-weight condition.

Table 4.7. Transition probabilities between conditional entropy values at 400 learning steps within one generation of the iterated learning model, for the five-weight initialization condition (LEFT) and random-weight initialization condition (RIGHT), normalized by row

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.40**	0.54
	0.1-0.2	0.12	**0.61**	0.28
	0.2-0.3	0.25	**0.36**	0.38
	0.3-0.4	0.17	**0.36**	0.46
	0.4-0.5	0.21	**0.41**	0.38
	0.5-0.6	0.22	**0.45**	0.32
	0.6-0.7	0.36	**0.47**	0.22
	0.7-0.8	0.40	**0.46**	0.14
	0.8-0.9	0.51	**0.43**	0.07
	0.9-1.0	0.62	**0.38**	–

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.63**	0.38
	0.1-0.2	0.28	**0.51**	0.21
	0.2-0.3	0.25	**0.32**	0.43
	0.3-0.4	0.22	**0.38**	0.39
	0.4-0.5	0.30	**0.37**	0.33
	0.5-0.6	0.35	**0.39**	0.26
	0.6-0.7	0.46	**0.38**	0.16
	0.7-0.8	0.52	**0.37**	0.11
	0.8-0.9	0.63	**0.33**	0.04
	0.9-1.0	0.73	**0.27**	–

With these general biases established – a strong bias towards consistent plural marking across lexical items and a more inconsistent bias towards more deterministic grammars – the next section details the results of modeling the experimental conditions from Smith & Wonnacott (2010), in which human participants showed a strong bias towards more deterministic grammars, and a more inconsistent bias towards consistent plural marking across lexical items.

4.3 Modeling Smith & Wonnacott (2010)

As described in §4.1, Smith & Wonnacott (2010) conducted an artificial language learning experiment in which human participants were arranged into iterated learning chains (10 chains of 5 participants each). The first member of each chain was trained on an artificial grammar in which, for each lexical item, the plural was marked 75% of the time with one of the plural markers (either fip or tay), and 25% of the time with the other marker. Each participant's productions during the test phase were used as the training data for the next participant in the chain. In the experimental results, the participants showed a consistent bias towards more deterministic grammars, with the conditional entropy over participants' productions cumulatively decreasing across the chains, and an inconsistent bias towards consistent plural marking across lexical items, with different chains showing different rates of use of the initial majority plural marker.

In order to model the experimental conditions in Smith & Wonnacott (2010), I ran simulations with the iterated learning model in which the initial teacher agent, in all cases, had a MaxEnt grammar which placed a probability of 75% on choosing the plural marker FIP for each lexical item, and a probability of 25% on TAY (the "75% FIP" grammar). Because of the symmetric nature of the constraint set, sets of simulations run with TAY as the preferred marker would be mirror images of the sets run with FIP as the preferred marker. I additionally tested the effects of the five-weight and random-weight conditions on new learner agents introduced into the generational chains.

§4.3.1 first discusses the results of learning within one generation, then §4.3.2 details the behavior that emerges when iterating over multiple generations of agents, and compares the model results to the experimental results in Smith & Wonnacott (2010). The results here are very similar to the results seen in the general exploration in §4.2. The iterated learning model here show a bias towards consistent plural marking across lexical items, away from exceptionality, as well as a bias away from variability that is less robust across modeling conditions. This is somewhat the opposite of the experimental results in Smith & Wonnacott (2010), as discussed in §4.3.3.

4.3.1 Iterated learning within one generation

In this section, I first detail the results of learning within one generation, from the initial teacher to the first generation of learners, to inform the later discussion of the results of iterating over multiple generations. Two sets of simulations were run: one where new learner agents were initialized under the five-weight condition, and one where they were initialized under the random-weight condition. In all simulations, the agents learned from their target grammar for 1,000 learning steps. 10,000 runs of each simulation were performed, with a learning rate of 0.1.

Table 4.8 summarizes the resulting distributions over pattern types and conditional entropy values, showing the results of both the five-weight and random-weight conditions, as well as repeating the sampled baseline distribution. Table 4.8 also shows the properties of the target 75% FIP grammar, which corresponds to the 4X0Y pattern type, and has a conditional entropy value of 0.81.

Neither the five-weight condition simulations nor the random-weight condition simulations fully fit the target grammar, despite seeming to reach a fairly stable distribution, as shown in Figures 4.8 and 4.9. They do, however, show a higher proportion of runs ending in 4X0Y patterns (over 85%) than in the baseline distribution (40%) or in the general exploration in §4.2 (around 75%). Additionally, the conditional entropy distributions show peaks at higher ranges, closer to the target value of 0.81, compared to the baseline distribution or the general exploration in §4.2.

Figure 4.8 plots the change in the proportion of runs in each pattern type across generations of the simulations. The five-weight condition is shown on the left, and the random-weight condition is shown on the right. Agents in a large majority of runs in both conditions acquire 4X0Y patterns. In the five-weight condition, this majority is attained instantly, and remains fairly stable across learning steps. In the random-weight condition, on the other hand, there is an initial decrease in the proportion of 4X0Y patterns, and decrease in 3X1Y and 2X2Y patterns, before the proportion of

Table 4.8. The distribution over pattern types and conditional entropy after one generation of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar

	Pattern Type		
	4X0Y	3X1Y	2X2Y
Sampled Baseline	0.40	0.40	0.20
Initial Target	1.00	0.00	0.00
Five-wght Cond.	0.86	0.05	0.01
Rand-wght Cond.	0.90	0.08	0.02

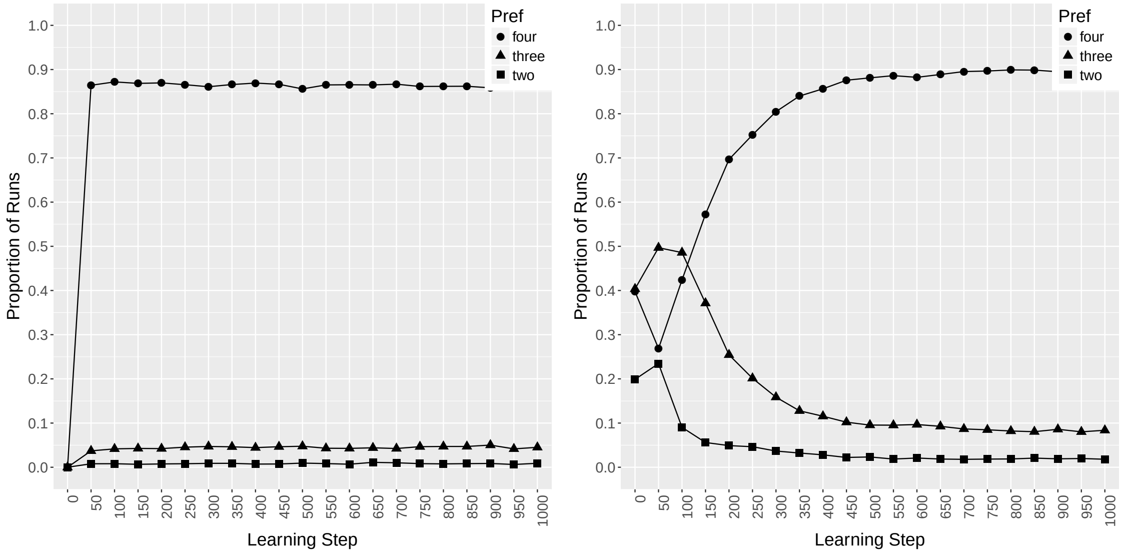
	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Initial Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Five-wght Cond.	0.00	0.00	0.00	0.01	0.02	0.06	0.28	0.19	0.29	0.14
Rand-wght Cond.	0.00	0.00	0.00	0.01	0.03	0.10	0.19	0.26	0.27	0.16

4X0Y patterns gradually increases and remains fairly stable at around 90% of runs. This initial disadvantage for 4X0Y patterns is due to the random initial weights for the learner agents. Agents whose initial grammars are very different from the target grammar will pass through various intermediate stages as they adjust their constraint weights to fit their training data.

Figure 4.9 plots the change in the distributions over conditional entropy values across generations. Agents in both initialization conditions are acquiring grammars with conditional entropy values closer to the target value of 0.81 than in either the baseline distribution or the general exploration in §4.2, showing the influence of the target grammar. However, the variance in both cases is quite large, with values ranging from 0.4-1.0, and remains quite stably so across the simulations.

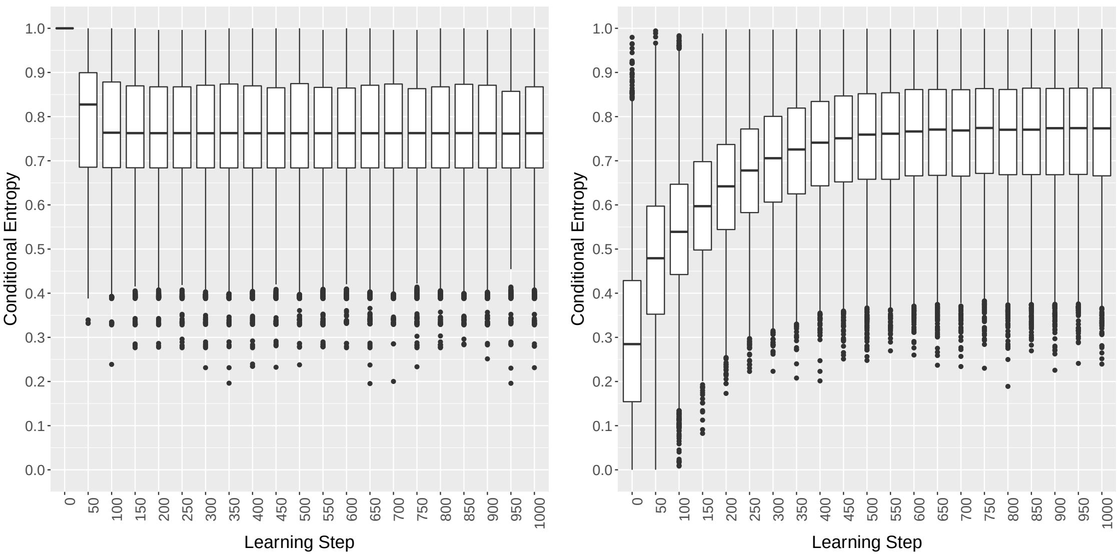
The noisiness in the agents' learning here may initially give a worrisome impression that the agents are not actually able to fit this target grammar. The source of this noise, however, can be attributed to various parameters of the simulations, including the learning rate and the sampling of the training data, and does not represent a

Figure 4.8. Plots showing the change in the distribution over pattern types across one generation of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (Left: five-weight condition, Right: random-weight condition)



failure of the model itself. Lowering the learning rate does decrease the noise in the agents' fit to the target grammar, as discussed in more detail in §4.4.2.

Figure 4.9. Plots showing the change in conditional entropy across one generation of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (Left: five-weight condition, Right: random-weight condition)



4.3.2 Iterated learning across generations

This section now turns to examining the results of iterating over multiple generations of agents, where the first learner agent in each generational chain is trained on the 75% FIP grammar, emulating the experimental conditions in Smith & Wonnacott (2010). Two sets of simulations were run, testing the five-weight and random-weight initialization conditions for learner agents. For both conditions, the learner agents learned from their teacher agent for 400 learning steps, and the simulation iterated over 50 generations. 10,000 runs of each simulation were performed, with a learning rate of 0.1.

Table 4.9 summarizes the resulting distributions over pattern types and conditional entropy values for both initialization conditions, as well as repeating the sampled baseline distribution and the properties of the initial 75% FIP grammar. Both initialization conditions show a bias towards consistent plural marking across lexical items, with over 75% of runs ending in 4X0Y patterns, compared to 40% of sampled sets of weights in the baseline distribution. This bias is not as strong, however, as the results of learning within one generation, and the results of both initialization conditions here are actually quite similar to the results of learning in the general case, described in §4.2, suggesting that the influence of the initial target grammar in the chain gets washed out over multiple generations of agents. The distributions over conditional entropy values here are also very similar to the results obtained in the general case, where only the random-weight condition shows a clear bias towards more deterministic grammars. In the random-weight condition, the distribution shows a higher peak at conditional entropy values less than 0.1 and between 0.1-0.2 (27% and 33%, respectively) compared to the sampled baseline distribution (18% and 14%, respectively). The five-weight condition, on the other hand, shows a higher peak at conditional entropy values between 0.1-0.2 (28%), but a lower proportion of values less than 0.1 (6%), and a higher proportion of values greater than 0.6.

Table 4.9. The distribution over pattern types and conditional entropy after 50 generations of learning with the iterated learning model in the plural marker typology, with a 75% fip initial target grammar

	Pattern Type		
	4x0y	3x1y	2x2y
Sampled Baseline	0.40	0.40	0.20
Initial Target	1.00	0.00	0.00
Five-wght Cond.	0.77	0.14	0.08
Rand-wght Cond.	0.79	0.15	0.09

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Initial Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Five-wght Cond.	0.06	0.28	0.17	0.10	0.09	0.09	0.10	0.07	0.03	0.01
Rand-wght Cond.	0.27	0.33	0.11	0.08	0.08	0.07	0.04	0.02	0.01	0.00

Figure 4.10 plots the change in the proportion of runs in each pattern type across generations of the simulations. The five-weight condition is shown on the left, and the random-weight condition is shown on the right. Both initialization conditions

show a very similar trajectory, where the proportion of runs in a 4x0y pattern starts out as a large majority, due to the influence of the initial 75% fip target grammar. This initial advantage for 4x0y patterns is diminished as the proportion of 3x1y and 2x2y patterns increases over the next 5-10 generations, before this trend is reversed, and the proportion of 4x0y patterns gradually increases across further generation, and the proportion of 3x1y and 2x2y patterns decreases.

Figure 4.10. Plots showing the change in the distribution over pattern types across 50 generations of the iterated learning model in the plural marker typology, with a 75% fip initial target grammar (Left: five-weight condition, Right: random-weight condition)

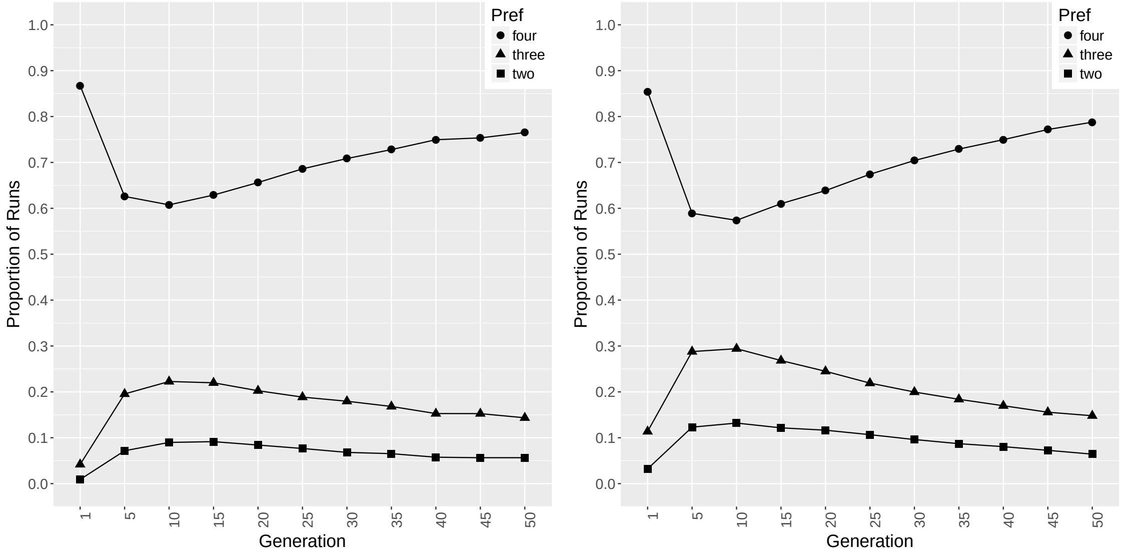
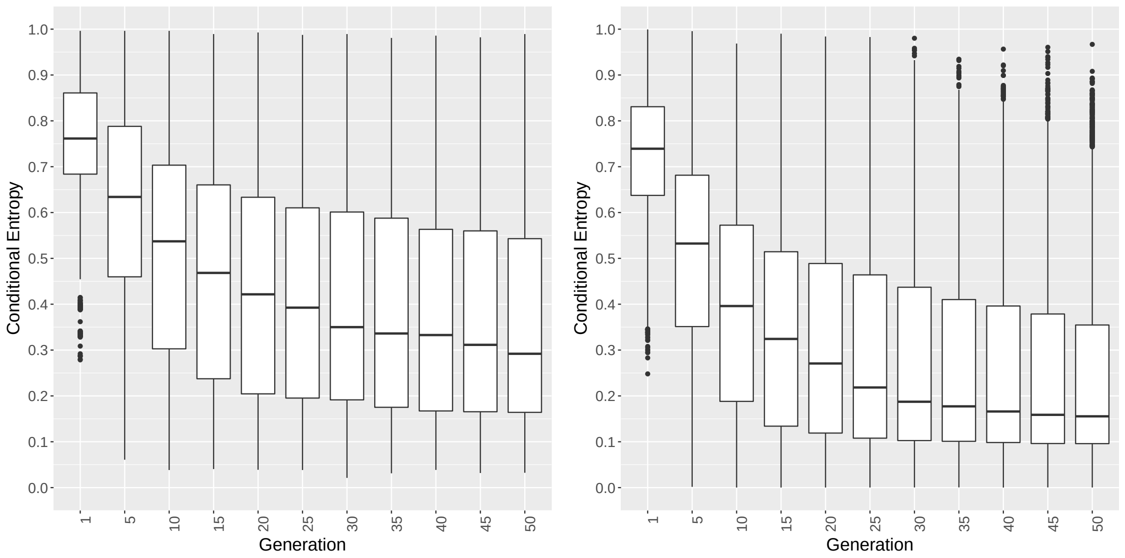


Figure 4.11 plots the change in the distribution over conditional entropy values across generations of the simulations. The five-weight condition is shown on the left, and the random-weight condition is shown on the right. Both initialization conditions, again, show a very similar trajectory, where the distribution gradually trends towards lower values, and thus more deterministic grammars, across generations. However, this effect is more striking in the random-weight condition, where both the mode of the distribution and the variance get smaller compared to the five-weight condition.

Figure 4.11. Plots showing the change in conditional entropy across 50 generations of the iterated learning model in the plural marker typology, with a 75% fip initial target grammar (Left: five-weight condition, Right: random-weight condition)



4.3.3 Discussion on modeling Smith & Wonnacott (2010)

§4.3 presented the results of modeling the experiment performed in Smith & Wonnacott (2010), in which human participants were tested on their ability to learn from an artificial grammar exhibiting free variation. Participants were arranged into iterated learning chains, where the first member of each chain was trained on an artificial grammar in which, for each lexical item, the plural was marked 75% of the time with one of the plural markers (either fip or tay), and 25% of the time with the other marker. Each participant's productions during the test phase were used as the training data for the next participant in the chain. The results of this experiment indicated a strong trend towards more deterministic patterns, but only an inconsistent trend towards preferring the same plural marker across all lexical items.

To model the experiment in Smith & Wonnacott (2010), iterated learning model simulations were performed, in which the initial target grammar in each generational chain placed a probability of 75% on the plural marker fip, for all lexical items. Two initialization conditions for new learner agents were tested: one in which agents were initialized with all constraint weights set to the arbitrary positive real number five (the five-weight condition), and one where all constraint weights were randomly sampled from a uniform distribution between 0.0-10.0 (the random-weight condition).

Both initialization conditions showed a consistent bias towards preferring the same plural marker across all lexical items when compared to the sampled baseline distribution, with higher proportions of runs ending in a 4x0y pattern. However, only the random-weight condition showed a clear bias towards more deterministic grammars, with higher proportions of runs ending with conditional entropy values closer to zero compared to the sampled baseline distribution. The five-weight condition showed mixed results, with a higher proportion of runs ending with conditional entropy values between 0.1-0.2, but a lower proportion ending with lower values, and a higher proportion ending in higher values, compared to the baseline distribution. These results are incredibly similar to the results of the general exploration in §4.2, in which the initial target grammars were randomly sampled, suggesting that the effect of the initial target grammar gets washed out as the simulation iterates over multiple generations. See §4.2.3 for a more detailed discussion of the source of these emergent biases in the iterated learning model.

The results of the iterated learning model simulations differ from the experimental results presented in Smith & Wonnacott (2010) in various ways. First, the agents in the iterated learning model showed a stronger bias towards consistent plural marking across lexical items than did the participants in the Smith & Wonnacott (2010) experiment. In the iterated learning model results, the final agent in over 75% of learning chains preferred the same plural marker on all lexical items, whereas only 50% of the final participants in the experimental learning chains (5/10 chains) showed a consistent preference for one plural marker. This can be seen in the left-hand graph in Figure 4.1, which shows the number of plurals the participants marked using the initial majority marker in each chain. Three participant chains marked 16/16 plurals with the initial majority marker, with one chain marking nearly 16/16, and one participant chain marked 0/16 plurals with the initial majority marker, thus using the other plural marker exclusively.

Secondly, the agents in the iterated learning model showed a weaker bias towards deterministic patterns than did the participants in the Smith & Wonnacott (2010) experiment. In the iterated learning model results, the final agents in the learning chains were concentrated at lower conditional entropy values, but were not fully deterministic, and showed a large amount of variance between runs. In the experimental results, on the other hand, the mean conditional entropy value across the final participants in the chains, and the range of the 95% confidence intervals, are highly concentrated at lower values of entropy, below 0.2.

The differences between the model results and the experimental results could be due to a number of factors that differ between the modalities. First, the sample size in Smith & Wonnacott (2010) is much smaller than in the iterated learning model results presented here. The experiment contained 10 diffusion chains, each 5 participants long, while the iterated learning model results contained 10,000 runs (chains of agents), each of which iterated over 50 generations. It is possible that the sample size in the experiment was too small, and that including more participant chains, or longer participant chains, would have revealed a stronger bias in favor of consistent plural marking.

It is also possible that the choice of using adult participants in Smith & Wonnacott (2010) provided a weaker bias source than children would have, and that the results of the iterated learning model, with these parameters, provide a more accurate model of child learning than adult learning. As the discussion of previous literature in §4.1 showed, children have shown stronger biases towards regularization than have adults, who have shown a greater tendency towards probability matching. However, this would only provide an explanation for the stronger bias towards 4x0y patterns in the iterated learning model than in the experiment results, and still leaves unexplained the weaker bias towards more deterministic patterns in the iterated learning model, compared to the human participants. The weaker bias towards more deterministic patterns in the iterated learning model would seem to indicate behavior closer to probability matching, compared to the behavior of the human participants in the experiment results.

Looking across the simulations presented in this chapter, the set yielding the results closest to the experimental results is actually the interactive learning model results under the random-weight condition. This set of simulations yielded the weakest bias towards 4X0Y patterns, though still stronger than in the experiment results, and the strongest bias towards more deterministic grammars, though still somewhat weaker than in the experiment results. This raises questions about the relationship between learning an artificial grammar in a laboratory setting and learning in a more naturalistic setting, and how the strategies that participants use in an experimental study fit into the range of learning dynamics simulated by the interactive and iterated learning models. However, it is also possible that simply performing some further experimentation with manipulating the iterated learning model parameters could provide results yielding a closer fit to the experimental results.

Clearly, more research must be done to better establish the biases that human learners hold towards variation and exceptionality on an individual level, the biases that emerge on a population level when examining the cross-linguistic typology, participants' learning strategies in artificial learning experiments, and the relationship between the learning model parameters and the behavior of human participants. These questions are left open for future research.

4.4 Parameter manipulations

This section discusses several questions and issues concerning model behavior and model parameters which emerged during the course of this chapter. §4.4.1 discusses the importance of the general constraints for the emergence of both the bias towards consistent plural marking. §4.4.2 discusses the effect of lowering the learning rate, and how the model biases rely on noisy learning. §4.4.3 discusses the interference that arises between setting a lower bound on constraint weights at zero and initializing the constraint weights at zero in this typology, motivating the use of the five-weight initialization condition for this chapter, rather than the zero-weight initialization condition used in previous chapters.

4.4.1 Importance of the general constraints

As observed in the model results in §4.2 and §4.3, both the interactive and the iterated learning models show a strong bias towards consistent plural marking across lexical items. This bias is due to the inclusion in the constraint set of the general constraints, which favor one plural marker across all lexical items. In order to demonstrate this reliance, iterated learning simulations were run to compare the results obtained when the general constraints are present in the constraint set, and the results obtained when they are excluded, and the agents are relying solely on the lexically-indexed constraints. In both cases, the initial learner agent in each chain was trained on data from a randomly sampled initial grammar (generated by randomly sampling constraint weights from a uniform distribution between 0-10), each learner agent was initialized with constraint weights at 5 and learned from its target grammar for 400 learning steps, the simulation iterated over 50 generations, and the learning rate was 0.1.

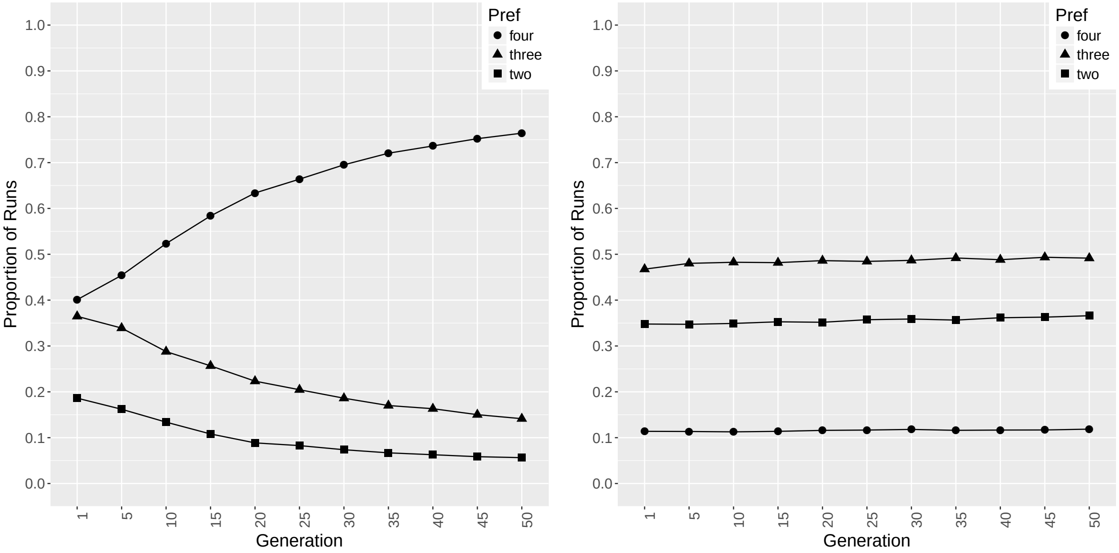
Table 4.10 summarizes the results of the iterated learning simulations performed with and without the general constraints, as well as giving sampled baseline distributions estimated with and without the inclusion of the general constraints. Comparing the sampled baseline distributions over pattern types, it can be seen that the baseline sampled without the general constraints contains a much smaller proportion of 4x0y patterns than does the baseline sampled with the general constraints, showing that the inclusion of the general constraints introduces many more possible weightings that yield a pattern in which the same plural marker is preferred across lexical items. In fact, the baseline distribution without the general constraints is simply the proportion of possible permutations of each pattern type in this typology; for example, referring to Table 4.2, there are 8 possible ways of making a 3x1y pattern, out of 16 possible patterns, yielding a proportion of 0.50. In examining the results after learning, it is clear that there is only a bias towards consistent plural marking when the general constraints are included; no such bias emerges when they are excluded. In the simulations run with the general constraints, 76% of the resulting grammars correspond to a 4x0y pattern, compared to 40% in the baseline sampled with the general constraints included. In the simulations run without the general constraints, only 12% of the resulting grammars correspond to a 4x0y pattern, and this distribution after learning hardly differs at all from the baseline sampled without the general constraints.

Table 4.10. The distribution over pattern types after 50 generations of the iterated learning model in the plural marker typology with random initial target grammars, comparing results obtained with the general constraints and with no general constraints

	Pattern Type		
	4x0y	3x1y	2x2y
Baseline With Gen.	0.40	0.40	0.20
Baseline No Gen.	0.13	0.50	0.37
Results With Gen.	0.76	0.14	0.06
Results No Gen.	0.12	0.49	0.37

The plots in Figure 4.12 take a closer look at the change in the distribution over pattern types across generations in the simulations. The results of the simulations performed with the general constraints included are shown in the left-hand graph, and the results with the general constraints excluded are shown in the right-hand graph. In the results with the general constraints, the proportion of 4X0Y patterns gradually increases to reach a sizable majority. In the results without the general constraints, proportion of each pattern type matches the proportion in the baseline distribution sampled without the general constraints throughout the course of the simulation, with no advantage for the 4X0Y patterns.

Figure 4.12. Plots showing the change in the distribution over pattern types across 50 generations of the iterated learning model in the plural marker typology, with random initial target grammars (Left: with general constraints, Right: with no general constraints)



Comparing the results of iterated learning model simulations performed with and without the general constraints preferring one plural marker in all contexts has revealed that the inclusion of these general constraints is crucial for the emergence of a bias towards consistent plural marking across lexical items. When the general constraints are excluded, the agents use solely the lexically-indexed constraints to represent their grammar. This effectively means that the probability distribution over outcomes for each lexical item is fitted individually, with no interaction between items, and thus no pressure to maintain any consistency across the lexicon. Even though the bias towards consistent plural marking observed in the results of the Smith & Wonnacott (2010) experiment was not as strong as that observed in the models' results, it does suggest a pressure towards more consistent patterns of the type enforced by the general constraints.

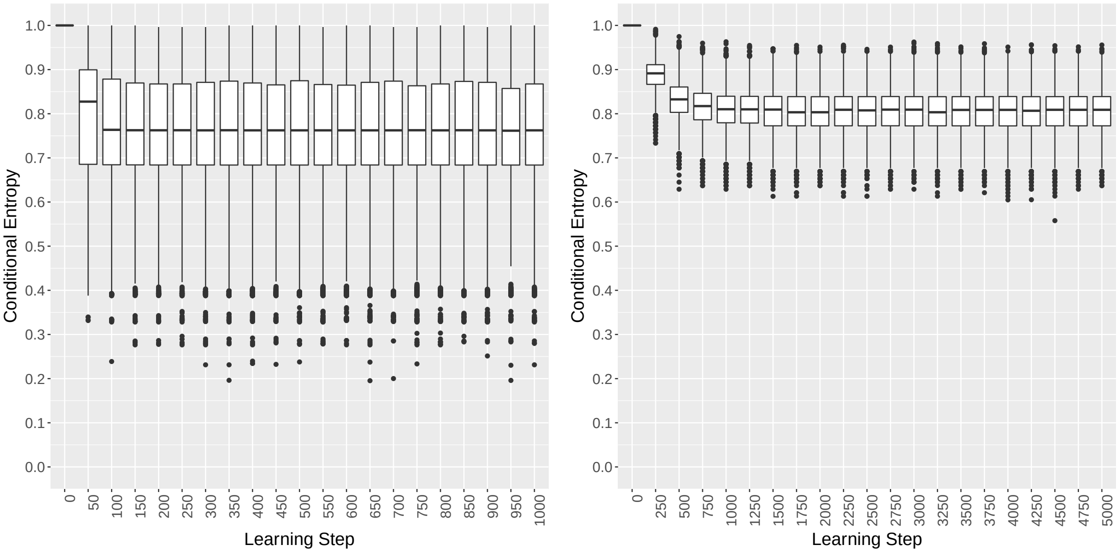
4.4.2 Consequences of reducing the learning rate

As observed in the results presented in §4.3.1, which examined learning of the 75% FIP grammar in a single generation, the learning agents are not always fully fitting the target grammar. Rather, learning is noisy, and the agents are ending in grammars within a particular range around the target. This can be seen most clearly in the distribution over conditional entropy values, as shown in the left-hand graph in Figure 4.13, which repeats the results of one generation of learning from the 75% FIP target grammar under the five-weight initialization condition. The target conditional entropy value for the 75% FIP grammar is 0.81, and although the mode of the distributions settle only slightly below that value, the variance of the distributions is quite large, and does not decrease appreciably as the number of learning steps increases.

This noisiness in the agents' learning may initially raise worries that the agents in these models are not actually able to fit their target grammars. However, this noise has its source in several of the model parameters, including the learning rate and the sampling of the training data, and doesn't represent a failure of the model itself. Decreasing the learning rate decreases the variance in conditional entropy across runs, leading to a closer average fit to the target grammar. This can be seen in the right-hand graph in Figure 4.13. This graph shows the results of one generation of learning from the 75% FIP target grammar under the five-weight initialization condition, for

5,000 learning steps, where the learning rate was lowered to 0.01, compared to the learning rate of 0.1 used elsewhere throughout this chapter. Under these conditions, the mode of the resulting distributions settles closer to the target value of 0.81, compared to the results with the learning rate of 0.1, and the variance of the distributions is much smaller.

Figure 4.13. Plots showing the change in the distribution over conditional entropy across one generation of the iterated learning model, with a 75% FIP initial target grammar (LEFT: learning rate 0.1, RIGHT: learning rate 0.01)



Although decreasing the learning rate in the simulations does increase the agents' average fit to the target grammar, this loss of noise in learning has the additional consequence of hindering the trend towards more deterministic grammars as the simulation is iterated over multiple generations. Under the random-weight initialization condition, the bias towards more deterministic grammars is simply weaker with the lower learning rate than with the higher, while under the five-weight initialization condition, the bias does not clearly emerge at all. In order to show this lack of bias in the five-weight condition, two sets of iterated learning model simulations are examined here, one in which the learning rate was 0.1, and agents learned from their target grammars for 400 learning steps, and one in which the learning rate was 0.01, and agents learned from their target grammars for 2,000 learning steps. In both sets, the initial target grammar was the 75% FIP grammar, new learner agents in the chain were initialized with constraint weights of five, and the simulations iterated over 50 generations of agents. The simulations with learning rate 0.1 here are the same as presented in §4.3.2.

Table 4.11 summarizes the results of the simulations run with a learning rate of 0.1 and a learning rate of 0.01, as well as repeating the sampled baseline distribution, and the properties of the initial 75% FIP target grammar. Comparing the resulting distributions over pattern types, it can be seen that the simulations run with the lower learning rate retain a higher proportion (88%) of 4X0Y patterns than do the simulations with the higher learning rate (77%). In both cases, these proportions are substantially higher than the sampled baseline distribution. Comparing the resulting distributions over conditional entropy values reveals a more significant difference between the learning rate conditions. The results of the simulations performed with a learning rate of 0.1 show a peak at lower conditional entropy values (between 0.1-0.2), showing a bias towards more deterministic grammars, while the simulations performed with a learning rate of 0.01 show a peak in the highest range of conditional entropy values (between 0.9-1.0), showing a trend towards more variable grammars. Both learning rate conditions, however, show a lot of variance, with a large number of runs ending in other ranges. The results of the 0.1 learning rate are more similar to the sampled baseline distribution, although the learning results contain a smaller proportion of conditional entropy values less than 0.1, while the results of the 0.01 learning rate are more similar to the initial 75% FIP grammar.

Figure 4.14 takes a closer look at the change in the distributions over pattern types across generations of the simulations. The simulations run with a learning rate of 0.1 are shown in the left-hand graph, while the simulations run with a learning rate of 0.01 are shown in the right-hand graph. The results of the higher learning rate

Table 4.11. The distribution over pattern types and conditional entropy after 50 generations of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar, comparing the results obtained with a learning rate of 0.1 and a learning rate of 0.01

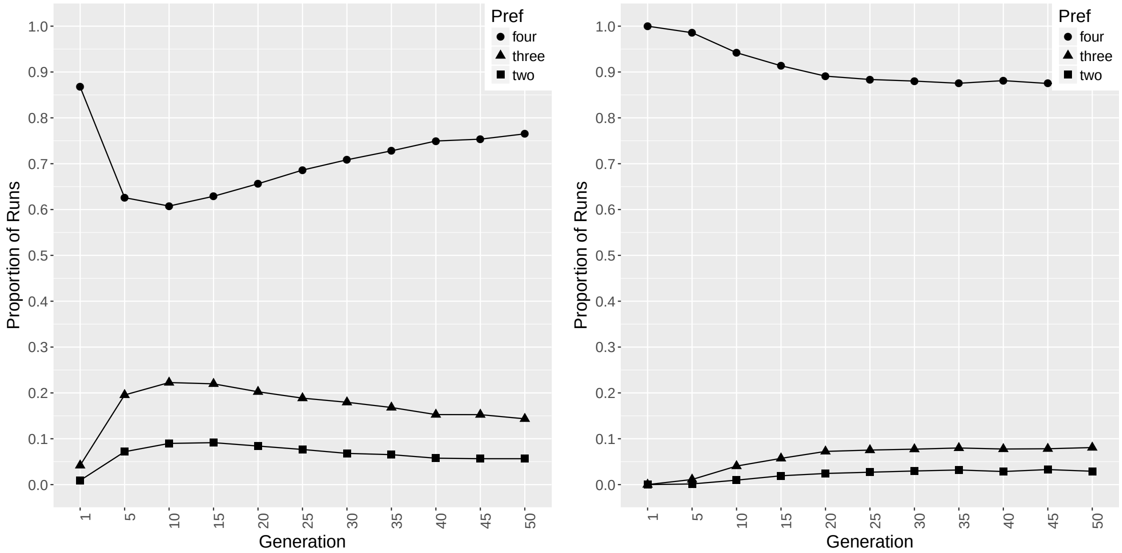
	Pattern Type		
	4X0Y	3X1Y	2X2Y
Sampled Baseline	0.40	0.40	0.20
Initial Target	1.00	0.00	0.00
LR 0.1	0.77	0.14	0.08
LR 0.01	0.88	0.08	0.03

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Initial Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
LR 0.1	0.06	0.28	0.17	0.10	0.09	0.09	0.10	0.07	0.03	0.01
LR 0.01	0.00	0.00	0.01	0.07	0.13	0.13	0.13	0.13	0.15	0.26

show an initial decrease in the proportion of 4X0Y patterns, which then gradually increases across generations, maintaining a sizable majority of runs in which the same plural marker is preferred on all lexical items. The results of the lower learning rate does not show this initial dip, but rather the proportion of 4X0Y patterns decreases gradually, while maintaining a sizable majority.

Figure 4.15 takes a closer look at the change in the distributions over conditional entropy values across generations of the simulations. The use of violin plots here allows for a better view of the skew in these distributions than would the box plots used elsewhere in this dissertation. The simulations run with a learning rate of 0.1 are shown in the left-hand graph, while the the simulations run with a learning rate of 0.01 are shown in the right-hand graph. The results of both learning rates show a large amount of variance across runs, as shown by the large spread of the distributions across conditional entropy values, though the lower learning rate produces somewhat less variance. With the learning rate of 0.1, however, there is a concentration of runs at lower conditional entropy values, with a long tail extending towards higher values,

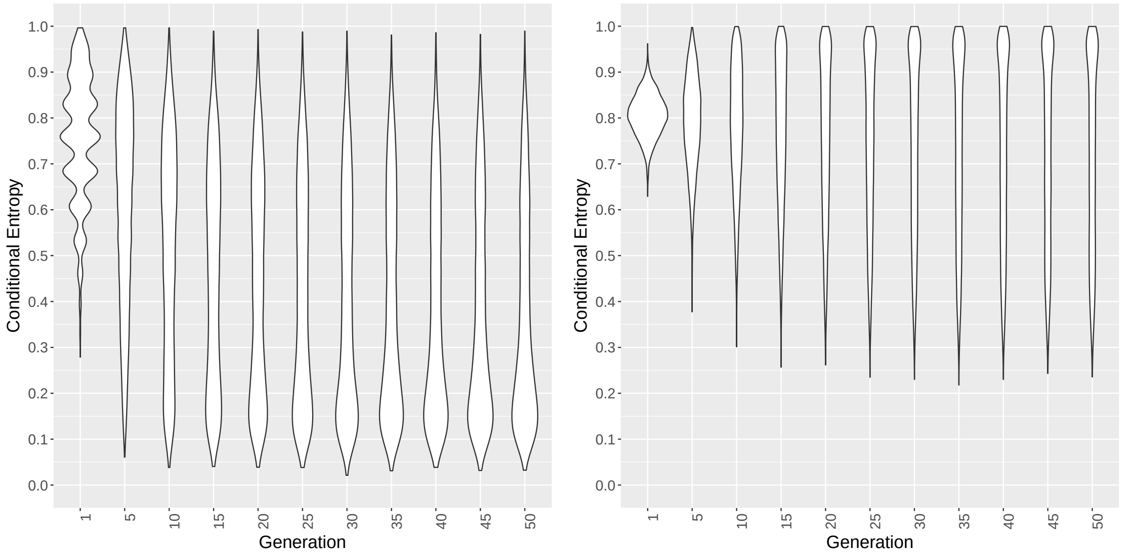
Figure 4.14. Plots showing the change in the distribution over pattern types across 50 generations of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (LEFT: learning rate 0.1, RIGHT: learning rate 0.01)



while with the learning rate of 0.01, there is a concentration at higher conditional entropy values, with a long tail extending towards lower values.

The explanation for the lack of a clear bias towards more deterministic grammars in the five-weight initialization condition for the lower learning rate of 0.01, compared to the higher learning rate of 0.1, stems from the reduced noise in agents' learning, and how that impacts the ways in which learning agents might mislearn their target grammar. As was seen in the results of learning within one generation, shown in Figure 4.13, the higher learning rate results in a larger degree of variance in learned conditional entropy values between runs than does the lower learning rate. This difference in the amount of variance between runs additionally results in different patterns of mislearning in learning agents in each condition. This can be seen in Table 4.12, which shows the transition probabilities between a target conditional entropy value in a given range, and whether the learned conditional entropy value was in a lower range, in the same range, or in a higher range. The results of learning for 400 learning steps with a learning rate of 0.1 are shown on the left, and the results

Figure 4.15. Plots showing the change in the distribution over conditional entropy across across 50 generations of the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (LEFT: learning rate 0.1, RIGHT: learning rate 0.01)



of learning for 2,000 learning steps with a learning rate of 0.01 are shown on the right. While the noisy learning produced by the lower learning rate allow agents in that condition to mislearn their target grammars in either direction, as either more deterministic or less deterministic grammars, the agents in the simulations run with a learning rate of 0.01 are vastly more likely to learn a grammar with a higher conditional entropy value than their target grammar. Reducing the noise in learning by lowering the learning rate, then, not only increases the agents' average fit to the target grammar, as observed in Figure 4.13, but also skews the direction of mislearning towards the initial state of the learning agents. For these simulations, new learner agents were initialized with all constraint weights at five, producing a completely variable grammar with a conditional entropy value of 1.0, and thus a skew towards more variable grammars across generations of agents.

Comparing the results of iterated learning simulations performed with a higher learning rate of 0.1 and a with a lower learning rate of 0.01 has revealed that decreasing the learning rate increases the learning agents' average fit to their target

Table 4.12. Transition probabilities between conditional entropy values after one generation of learning with the iterated learning model under the five-weight initialization condition, for (LEFT) learning rate 0.1 at 400 learning steps and (RIGHT) learning rate 0.01 at 2000 learning steps, normalized by row

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.40**	0.54
	0.1-0.2	0.12	**0.61**	0.28
	0.2-0.3	0.25	**0.36**	0.38
	0.3-0.4	0.17	**0.36**	0.46
	0.4-0.5	0.21	**0.41**	0.38
	0.5-0.6	0.22	**0.45**	0.32
	0.6-0.7	0.36	**0.47**	0.22
	0.7-0.8	0.40	**0.46**	0.14
	0.8-0.9	0.51	**0.43**	0.07
	0.9-1.0	0.62	**0.38**	–

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.00**	1.00
	0.1-0.2	0.00	**0.09**	0.91
	0.2-0.3	0.00	**0.06**	0.94
	0.3-0.4	0.00	**0.05**	0.95
	0.4-0.5	0.00	**0.04**	0.96
	0.5-0.6	0.00	**0.06**	0.94
	0.6-0.7	0.01	**0.13**	0.87
	0.7-0.8	0.01	**0.37**	0.62
	0.8-0.9	0.02	**0.89**	0.09
	0.9-1.0	0.13	**0.88**	–

grammar. However, reducing noise in learning by lowering the learning rate additionally reduces the amount of change that occurs between learning steps or generations of a simulation, which in turn impacts whether and how learning biases emerge over the course of the simulations. In the case of the iterated learning simulations presented in this section, which were performed under the five-weight initialization condition for new learner agents with the 75% FIP initial target grammar, the lower learning rate produced an increased bias towards 4X0Y patterns, showing increased faithfulness to their target pattern, but no visible bias towards more deterministic grammars, due to the influence of the agents' initial fully variable grammars. Noisy learning, then, is necessary in these learning models in order to observe the emergence of learning biases within the timescales of the simulations. Noisy learning in humans is also likely to be crucial in order to observe any linguistic change across time, however, change in natural languages occurs over much larger timescales, and so the amount of noise attributed to any individual is likely to be quite small.

4.4.3 Comparing the five-weight and zero-weight initialization conditions

In Chapter 2 and Chapter 3, two initialization conditions were tested for new learner agents, one in which each agent was initialized with each constraint weight randomly sampled from a uniform distribution with range 0-10 (the random-weight condition), and one in which each agent was initialized with constraint weights set at zero (the zero-weight condition). This chapter also tested the random-weight condition; however, the zero-weight condition was substituted for one in which each agent was initialized with constraint weights set to the same positive real number, in this case 5 (the five-weight condition). The substitution of the five-weight initialization condition for the zero-weight condition is possible because of the symmetric nature of the constraint set used for the plural marker typology, which can be observed in Tableau 4.1. Under the zero-weight initialization condition, in any typology, each competing output candidate will receive equal probability, yielding an initial state in which learner agents have no bias towards any output. In the plural marker typology, the violation profiles of each output candidate are similar, each incurring one violation each of two constraints, meaning that equal probability will be assigned to all competing candidates so long as each constraint has the same weight.

The advantage for the five-weight condition over the zero-weight condition, for this typology, is that it produces an equivalent initial state, where all competing output candidates have equal probability, while avoiding any interaction with the restriction that constraint weights be non-negative real numbers. Under the zero-weight initialization condition, as learner agents exchange data with the other agent in the simulation, they gradually adjust their constraint weights away from zero in response to any errors they produce. The restriction on possible values of constraint weights is implemented in learning as a lower bound on constraint weights such that they cannot fall below zero. If a learning update would result in a constraint weight below zero, the weight of that constraint is simply set to zero, meaning that the full intent of the grammar update is not effected. In the cases of the typologies in Chapter 2 and Chapter 3, the violation profiles of competing output candidates were not fully symmetric, producing a set of possible grammar updates with sufficient variation to obscure the effects of clipping the constraint weights at zero. In the case of the plural marker typology, however, the violation profiles of the competing output candidates are fully symmetric, and this symmetry appears to interfere with learning in these simulations when the constraint weights are initialized with zero, and prevented from falling below zero. Either removing the lower bound on constraint weights, or initializing all constraint weights at a non-zero positive number, mitigates the effects of this interference.

In order to show the differences between the zero-weight condition and the five-weight condition, two sets of iterated learning simulations, one with each initialization condition, are compared below. In both cases, the initial target grammar in the chain was the 75% FIP grammar, the learner agents learned from their target for 400 learning steps, the simulation iterated over 50 generations, and the learning rate was 0.1. Table 4.13 summarizes the results of the five-weight and zero-weight initialization conditions, repeating also the sampled baseline distribution and the properties of the 75% FIP initial target grammar. Both initialization conditions show a bias towards consistent plural marking across lexical items, with the final agent in the chain ending in a 4X0Y pattern in over 75% of runs, compared to just 40% in the sampled baseline distribution. The primary difference between the results of these conditions lies in the distribution over conditional entropy values. The five-weight condition shows a tendency towards more deterministic grammars, as evidenced by the peak in the distribution at lower conditional entropy values (between 0.1-0.2), while the results of the zero-weight condition show a fairly even spread over conditional entropy values greater than 0.2, with no clear tendency towards more or less deterministic grammars.

Table 4.13. The distribution over pattern types and conditional entropy after 50 generations of learning with the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar, comparing the five-weight and zero-weight initialization conditions

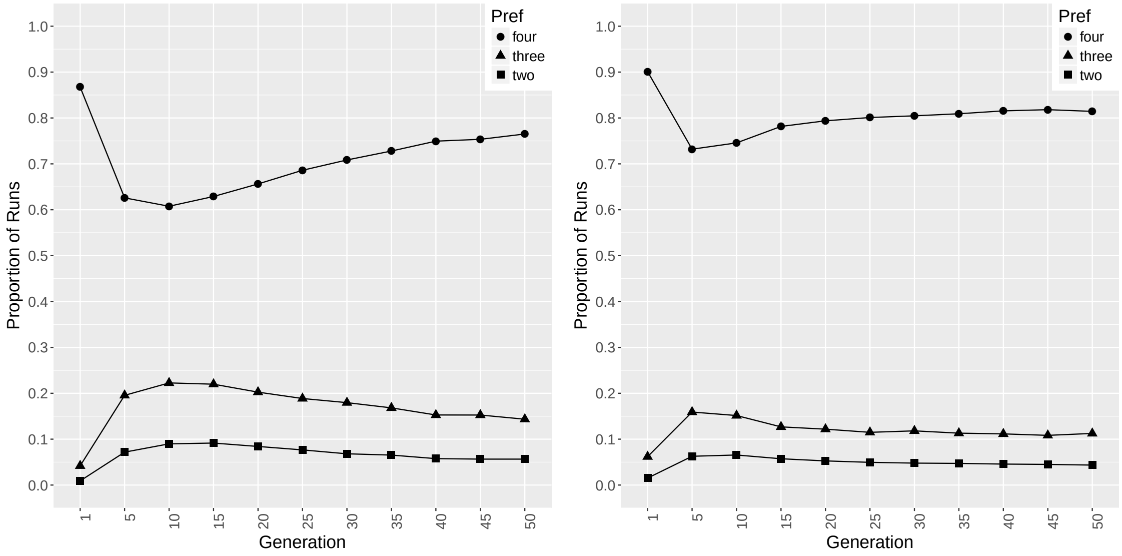
	Pattern Type		
	4X0Y	3X1Y	2X2Y
Sampled Baseline	0.40	0.40	0.20
Initial Target	1.00	0.00	0.00
Five-wght Cond.	0.77	0.14	0.08
Zero-wght Cond.	0.81	0.11	0.04

	Conditional Entropy									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Sampled Baseline	0.18	0.14	0.20	0.18	0.14	0.09	0.04	0.02	0.00	0.00
Initial Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Five-wght Cond.	0.06	0.28	0.17	0.10	0.09	0.09	0.10	0.07	0.03	0.01
Zero-wght Cond.	0.00	0.02	0.12	0.14	0.13	0.11	0.11	0.12	0.15	0.10

The plots in Figure 4.16 take a closer look at the change in the distribution over pattern types across generations in the simulations. The five-weight condition is shown on the left, and the zero-weight condition is shown on the right. Both conditions show a fairly similar timeline, with an initial dip in the proportion of 4X0Y patterns, which then gradually increases across generations. This effect, however, is weaker in the zero-weight condition than in the five-weight condition.

The plots in Figure 4.17 take a closer look at the change in the distribution over conditional entropy values across generations in the simulations. Again, the use of violin plots here allows for a better view of the skew in these distributions than would the box plots used elsewhere in this dissertation. In the five-weight condition, the peak of the distribution gradually shifts towards lower conditional entropy values across generations, showing a tendency towards more deterministic grammars, though the distribution does retain a long tail. In the zero-weight condition, on the other hand, the distribution develops no clear peak at either more or less deterministic grammars, but rather seems to spread fairly evenly across conditional entropy values.

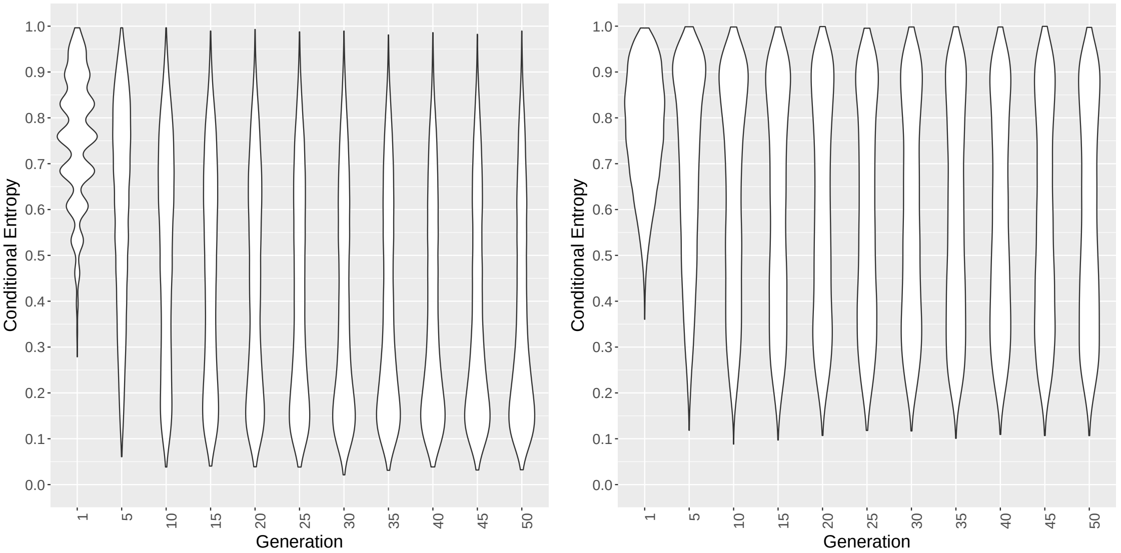
Figure 4.16. Plots showing the change in the distribution over pattern types across 50 generations of learning with the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (Left: five-weight condition, Right: zero-weight condition)



The difference between the zero-weight and five-weight conditions seems to be most pronounced when learning more deterministic target grammars, as can be seen in examining Table 4.14, which shows the transition probabilities between a target conditional entropy value in a given range, and whether the learned value was in a lower range, in the same range, or in a higher range. The results of the five-weight condition are shown in the left-hand table, and the results of the zero-weight condition are shown in the right-hand table. In the zero-weight condition, only 2% of agents learning a target grammar in the lowest range of conditional entropy values ended with a grammar in the same range, compared to 40% of agents in the five-weight condition. Additionally, the zero-weight condition generally shows a greater tendency towards mislearning towards higher conditional entropy values across higher ranges of target values compared to the five-weight condition.

Comparing the results of the five-weight initialization condition and the zero-weight initialization condition for the plural marker typology has shown that, even though both conditions produce similar initial states, where all competing output

Figure 4.17. Plots showing the change in conditional entropy across 50 generations of learning with the iterated learning model in the plural marker typology, with a 75% FIP initial target grammar (Left: five-weight condition, Right: zero-weight condition)



candidates have equal probability, the symmetric nature of the constraint set in this typology combined with the lower bound on constraint weights at zero causes interference in learning when the agents' initial constraint weights are set at zero. It is not immediately clear why the combination of the zero-weight initialization condition and the lower bound on weights at zero should have a noticeably more pronounced effect on the results of simulations performed over a typology with a symmetric constraint set than one without, and this question warrants further investigation in future work.

Table 4.14. Transition probabilities between conditional entropy values within one generation of the iterated learning model, at 400 learning steps, for the five-weight condition (LEFT) and the zero-weight condition (RIGHT), normalized by row

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.40**	0.54
	0.1-0.2	0.12	**0.61**	0.28
	0.2-0.3	0.25	**0.36**	0.38
	0.3-0.4	0.17	**0.36**	0.46
	0.4-0.5	0.21	**0.41**	0.38
	0.5-0.6	0.22	**0.45**	0.32
	0.6-0.7	0.36	**0.47**	0.22
	0.7-0.8	0.40	**0.46**	0.14
	0.8-0.9	0.51	**0.43**	0.07
	0.9-1.0	0.62	**0.38**	–

		Learned		
		Lower	Same	Greater
Target	0.0-0.1	–	**0.02**	0.98
	0.1-0.2	0.00	**0.32**	0.68
	0.2-0.3	0.10	**0.41**	0.49
	0.3-0.4	0.23	**0.32**	0.45
	0.4-0.5	0.25	**0.26**	0.49
	0.5-0.6	0.26	**0.26**	0.48
	0.6-0.7	0.25	**0.29**	0.46
	0.7-0.8	0.25	**0.38**	0.37
	0.8-0.9	0.29	**0.52**	0.20
	0.9-1.0	0.46	**0.54**	–

4.5 Discussion

The aim of this chapter was to investigate, in further detail, the bias towards more deterministic grammars shown by the interactive and iterated learning models in the results of simulations run with the simple test typology in Chapter 2, and again with the palatalization typology in Chapter 3. This chapter additionally sought to disambiguate between variation, where multiple output realizations are possible for a given input, and exceptionality, where different morphemes may arbitrarily display different behavior. In the absence of a ready typology from which it would be possible to derive estimations of the frequency of patterns of variation and patterns of exceptionality, relative to fully regular or deterministic patterns, the point of comparison for the behavior of the learning models instead relies on the body of previous literature which investigates the biases shown by human participants in artificial grammar learning experiments, overviewed in §4.1. The experimental results across this body of literature demonstrate that both adults and children can learn patterns of variation and exceptionality, but show a bias towards regularizing variation observed in their input. This bias away from variation is stronger in children than in adults.

The particular typology test case used throughout this chapter used the artificial grammar from the experiment in Smith & Wonnacott (2010), which tested adult participants on their ability to learn from an artificial grammar exhibiting free variation, and traced the evolution of the grammar as it was transmitted across participants in iterated learning chains (see §4.1.1.1 for more details). In the results of this experiment, participants showed a strong tendency towards producing patterns which were more deterministic than their training data, but did not show a strong bias towards using a single plural marker consistently across lexical items. Rather, about 50% of participant chains ended in a grammar which used one plural marker exclusively, or nearly exclusively, while other chains ended in patterns of exceptionality, in which the each plural marker was indexed to a subset of lexical items.

In order to model the results of Smith & Wonnacott (2010), I devised a set of constraints which included general constraints, which favored one of the plural markers in all contexts, as well as lexically-indexed constraints which favored one of the plural markers for a given lexical item. Although it is possible to represent patterns of consistent plural marker use across lexical items using the lexically-indexed constraints alone, the general constraints are necessary for generating a bias towards consistent plural marking, as discussed in §4.4.1.

In §4.2, I discussed the results of testing both the interactive and iterated learning models on this typological space in the general case, testing two initialization conditions: one where learner agents were initialized with all constraint weights set to five (the five-weight condition), and one where learner agents were initialized with each constraint weight randomly sampled from a uniform distribution with range 0.0-10.0. The five-weight condition was used in this chapter, rather than the zero-weight condition used in Chapter 2 and Chapter 3, due to interference caused by the symmetric nature of the constraint set in the plural marker typology and the lower bound at zero on constraint weights, as discussed in §4.4.3. In the simulations with the interactive learning model, both agents were initialized with the same constraint weights, while in the iterated learning model, the initial target grammars for each chain were generated by randomly sampling constraint weights, and the five-weight and random-weight initialization conditions applied to each new learner agent introduced into the chain. In §4.3, I discussed the results of modeling the Smith & Wonnacott (2010) experiment more explicitly, using the iterated learning model. In this case, the initial teacher agent in each chain held a grammar which placed a probability of 75% on the plural marker FIP for each lexical item, and new learner agents in the chain were tested using the five-weight and random-weight initialization conditions.

The results of the learning models, in all cases, showed a strong bias towards consistent plural marking across lexical items, as well as a less consistent bias towards more deterministic patterns. The bias for deterministic patterns was stronger for the interactive learning model than for the iterated learning model, and within the iterated learning model results, was stronger in the random-weight condition than in the five-weight condition. As discussed in §4.3.3, the differences between the experimental results, where participants showed a stronger bias towards more deterministic patterns and a weaker bias towards consistent plural marking, and the modeling results, in which agents showed a stronger bias towards consistent plural marking and a weaker bias towards more deterministic grammars, could be due to various sources, including the differing sample sizes between the experiment and the learning models, the choice of adult rather than child participants, the choice of model parameters, and uncertainty about how participants relate to the experiment task. These open questions are left for future work.

CHAPTER 5

CONCLUSIONS

5.1 Review

In this dissertation, I show how a theory of grammatical representations and a theory of learning can be combined to generate gradient typologies, predicting not only which patterns are expected to exist, but also their relative frequencies. Because not all natural language patterns are equally frequent cross-linguistically, this approach forms a more complete model of typology than previous approaches focusing on categorical typologies, which only make a binary prediction between patterns that are expected to exist and those that are not, based on the representational capacities of the assumed theory of grammar. The addition of a learning model component allows for the influence of learning biases in generating typological predictions: patterns which are learned more easily are predicted to be more typologically frequent than those which are learned more slowly.

The specific implementation of this system in this dissertation, generating gradient phonological typologies, uses Maximum Entropy grammar (Goldwater & Johnson, 2003) as the theory of grammar, combined with two agent-based learning models, the iterated learning model and interactive learning model, each of which mimics a type of learning dynamic observed in natural language acquisition. This system is tested on specific examples, showing how the models generate a bias away from patterns which rely on cumulative constraint interaction, and a bias away from variable patterns, both of which match up with observed trends in natural language typology. Future work can explore predictions for other typological trends, and explore the possibility of building a more complete learning model which combines the dynamics implemented in the iterated and interactive models.

5.1.1 Methodology

In Chapter 1, I detailed the methods used in this dissertation to generate gradient typological predictions in phonology through the combination of a theory of grammar and a theory of learning. In §1.3, I discussed the properties of the theory of grammar assumed in this work, Maximum Entropy grammar (MaxEnt; Goldwater & Johnson, 2003). MaxEnt is a probabilistic extension of Harmonic Grammar (HG; Legendre et al., 2006), a constraint-based theory of phonological grammar in which constraints are weighted, rather than ranked as in classic Optimality Theory (OT; Prince & Smolensky, 1993/2004). In MaxEnt, these constraint weights are used to define a probability distribution over outputs, allowing the grammar to represent patterns of variation between possible outputs.

In §1.4 I discussed the two agent-based learning models used in this work, the iterated learning model (§1.4.1) and the interactive learning model (§1.4.2), each of which aims to model different dynamics of interaction between agents in language acquisition. The iterated learning model (see e.g. Kirby & Hurford, 2002; Kirby et al., 2014) is a model of the vertical transmission of a language from an experienced user to an inexperienced user, for example from a parent to a child. Once the inexperienced user becomes an experienced user, they in turn transmit their language to a new inexperienced user, and this process repeats over multiple cycles, analogous to a language being passed from parents to children across generations. The pattern of interaction between agents in this model is unilateral in each generation: the learner agent is trained on data provided by a teacher agent, whose grammar is fixed.

The interactive learning model (see e.g. Pater, 2012), on the other hand, is a model of parallel, reciprocal communication between language users. Each agent in the population is learning from data produced by other agents, a dynamic which has analogues in linguistic accommodation and the spread of new linguistic conventions through a population. The pattern of interaction between agents in this model is bilateral (or multilateral, depending on the size of the population): each agent is both producing data for other agents to learn from, and learning from data produced by other agents.

In order to generate gradient typological predictions from the combination of MaxEnt with either the iterated or interactive learning model, multiple runs of each learning simulation were performed, as described in §1.5. In a given run of a simulation, each agent is initialized with a set of starting constraint weights, and learns from training data sampled from the other agent, as described in Algorithm 1 for the iterated learning model and Algorithm 2 for the interactive learning model, yielding one possible learning outcome. The distribution over outcomes across multiple runs of the simulation is taken as the predicted gradient typology.

As described in §1.3.3, for the purposes of this work, the resulting typology is defined as the combination of the distribution over pattern types, and the distribution over probability distributions. In Chapter 4, the distribution over probabilities is measured by conditional entropy (see equation in (1.5)). The pattern type metric identifies the set of highest probability output candidates, one for each input, categorizing the MaxEnt grammar according to the type of pattern, but abstracting over the probability distribution. The probability distribution metric serves as a measure of how evenly the probability mass is distributed among candidates, categorizing the MaxEnt grammar according to how variable it is, but abstracting over the pattern type.

The effects of combining MaxEnt with the learning models are evaluated by comparing the gradient typology generated as described above to a baseline distribution generated by sampling random constraint weights, to generate random MaxEnt grammars, without applying learning. Any differences between the predicted typology with learning and the baseline distribution can be attributed to the effects of learning biases which emerge through the influence of the learning model.

5.1.2 Model illustration: the simple test typology

In Chapter 2, I illustrate how this system of generating gradient typological predictions can be implemented, using a simplified, abstract typological space consisting of two constraints, two input forms, and two possible output candidates for each input. There are three possible pattern types in this typological space, one of which (the AC pattern) is only possible as a gang effect: a product of cumulative constraint interaction in which multiple violations of a lower-weighted constraint outweigh a single violation of a higher-weighted constraint. The results of simulations performed with both the iterated and interactive learning models show an emergent learning bias away from the gang effect AC pattern, reducing its predicted frequency relative to the baseline distribution, as well as an emergent learning bias away from variable grammars, such that the predicted typology with learning shows a stronger skew towards more deterministic patterns relative to the baseline distribution.

Closer examination of the results of the learning model simulations reveals that these biases – away from gang effects and away from variation – are linked due to the structure of the grammatical space. The weight of each constraint can be conceptualized as an axis in a multi-dimensional space – in the simple test case, this space is a two-dimensional plane – and any given possible grammar can be plotted as a point in this space. Pattern types can then be mapped as regions of the space which contain all grammars yielding that pattern type, and the borders between different pattern type regions correspond to areas with more variable grammars. These variable areas around the borders between pattern types are also areas of relative instability, as agents are more likely to move out of these areas than they are to move into them, resulting in a bias away from variable grammars.

This bias away from variable patterns is linked to the bias away from gang effects, as the nature of gang effect patterns means that regions corresponding to gang effects are bordered on all sides by other patterns, whereas other patterns might have a border along an axis. Borders with other patterns yield areas of variation, while borders along an axis do not, meaning that a greater proportion of the regions corresponding to gang effects also produce more variable grammars, compared to other pattern types. This combination of factors results in a learning disadvantage for gang effect patterns, and thus the bias away from these patterns observed in the predicted gradient typology.

5.1.3 Gang effect bias: the contrast types typology

In Chapter 3, I investigate the models' emergent bias away from gang effects further, showing how it is consistent with trends observed in natural language typology. I use a typology of contrast types as an empirical test case, which covers possible patterns of contrast between the fricatives /s/ and /ʃ/, in the contexts before the high front vowel /i/ or before any other vowel, represented with the label /a/. Modeling this typology using a standard set of three constraints – one banning /s/ before /i/, one banning /ʃ/ in any context, and one demanding faithfulness to the underlying representation – yields five possible patterns of contrast, one of which, the Elsewhere Neutralization pattern, is the result of a gang effect.

According to the estimated empirical frequencies of the contrast patterns in this typology, the gang effect Elsewhere Neutralization pattern is the rarest by far, being observed in only one natural language. In the results of simulations performed with the interactive and iterated learning models, both models produced a bias against the gang effect Elsewhere Neutralization pattern relative to a sampled baseline distribution, consistent with the empirical rarity of this pattern. Although the predicted typological distributions produced by the models did not perfectly line up with the observed empirical distribution, this could be due to a number of factors, such as an incomplete picture of the empirical probabilities of these patterns, or an insufficient incorporation of substantive biases into the learning models. These further considerations are left for future work.

5.1.4 Variation bias: the plural marker typology

In Chapter 4, I investigate the models' emergent bias away from variation further, showing how it is consistent with trends observed in human learning behavior. Because of the lack of an empirical estimation of the typological frequencies of patterns of variation compared to deterministic patterns, I used an empirical test case based on experimental findings with artificial grammar learning in human participants (Smith & Wonnacott, 2010). This test case was chosen additionally to disambiguate between variable behavior within a lexical item, i.e. where one input can be associated with multiple possible outputs, and variable behavior across lexical items, i.e. where different lexical items might display different behavior, though that behavior may be deterministic for each lexical item. The first type of variable behavior is labeled "variation", and the second type "exceptionality".

The empirical findings across an array of experimental studies suggest that while human participants in the lab are able to learn patterns of variation and exceptionality, patterns of variation are more difficult to learn than patterns of exceptionality. The results of simulations performed with the interactive and iterated learning models is consistent with the observed bias away from variable patterns, however, the models display a bias against patterns of exceptionality that is stronger than is suggested by the human behavior in the experimental results. Future work can address lingering questions about the relationship between the models' behavior and the behavior of the human participants, as well as questions about the strength of an exceptionality bias in human learning.

5.2 Next Steps

5.2.1 Typology of variation

One important next step is establishing an empirical estimation of the typology of variation in phonology. As discussed in Chapter 4, there is currently no systematic collection of empirical data on the frequencies of patterns of variation relative to deterministic patterns. This is likely due, at least in part, to the historical focus on deterministic patterns in typological surveys and studies. Given the rise in interest in accounting for patterns of variation in phonology (see Coetzee & Pater, 2011, for an overview), constructing such a typology would aid in understanding both the range of possible patterns of variation, and their relationship to more deterministic patterns. Though the host of available language grammars and research on variable patterns provides a rich source of attested patterns, collecting and compiling the necessary information into a usable database will take time, and will depend on the level of detail reported by the sources about the nature and extent of variation.

5.2.2 Model parameters

The major parameters to be set, for both the interactive and the iterated learning models, are the learning rate, the initial constraint weights for the agents, the number of learning steps allowed to each learner agent, and for the iterated learning model, the number of generations of agents. Throughout the dissertation, relatively little space has been dedicated to considerations of the effects of varying these parameters, which were set to arbitrary values which produced results which trended in the directions observed in the empirical data. Rather, the main focus in each case was kept on the presentation of the emergent biases produced by the interactive and iterated learning models, which were robust even without fine-tuning the parameter settings. However, as discussed in most detail in §4.4, the behavior of these models does vary with the parameter settings, and there are circumstances under which the biases away from gang effects and away from variation might be weakened or eliminated. Future work can provide a more thorough investigation into the effects of varying the parameter values, how robust these biases remain across conditions, and how these parameters might relate to human cognitive abilities.

5.2.3 Extension to other typologies

For the sake of simplicity, and enabling a more straightforward analysis of the results of the learning models, each of the typological examples explored in this dissertation have been restricted to cases where each input has only two competing possible output forms. While this restriction is useful for reducing the scope of the typologies considered, it is unnaturalistic, as different languages might employ different strategies to avoid unwanted phonological structures. Future work can test whether the emergent biases observed in the results of the iterated and interactive learning models in this dissertation hold when this system is extended to more complex typologies with larger candidate sets and constraint sets. In addition, the biases away from gang effects and away from variation found in this dissertation represent only a small subset of the possible typological trends that could be explored using this methodology. Future work might examine other typological trends, such as the generalization that no language disallows open syllables (Jakobson, 1962; Prince & Smolensky, 1993/2004).

5.2.4 Extending model functionality

In addition to extending the iterated and interactive learning models to other typological test cases, another area for future work is to investigate elaborations to the learning models themselves. In the simulations performed in this dissertation, there were only two agents in the population at a time, and the learner agents were trained on input-output pairs. Both of these assumptions simplify the learning problem and the subsequent analysis of the performance of the learning models, but they are unrealistic. Human language learners learn from and interact with a larger population of language users over the course of acquisition, and so one immediate next step for elaborating the learning models is to test the effects of varying population size (as was done in Reali et al., 2014), and varying the degree of connectedness between agents in the population. Varying the degree of connectedness between agents in the population could take several forms, including allowing agents to interact with any other agent equally, or restricting their interactions to a subset of the other agents (akin to a social network), or giving agents different probabilities of interacting with each other agent in the population (see Daland et al., 2007, for an example of using connection probabilities between agents). In addition, human language learners do not have direct access to the underlying meaning behind an observed utterance, whereas the agents in this dissertation had access to the input that produced the observed output. Humans must instead infer the underlying meaning of the observed utterances, which is a kind of hidden structure learning problem (see e.g. Tesar, 1998; Jarosz, 2013; Boersma & Pater, 2016, for various approaches to hidden structure problems). Another next step, then, is to incorporate an extra inference step in the learning process, in which agents must map the observed output to an available input.

5.3 Discussion

5.3.1 Emergent and innate biases

In this dissertation, I have shown and discussed two biases produced by the iterated and interactive learning models – one away from gang effect patterns, and one away from variation in the grammar – both of which echo trends observed in natural language typology. Importantly, these biases emerge over the course of learning, through the interactions between agents in the models, and do not need to be explicitly encoded into either the structure of the grammar or the learning algorithm. In demonstrating these emergent biases which result from the combination of a MaxEnt grammar and the iterated and interactive learning models, this dissertation adds to broader research questions in linguistics and phonology.

One such question is the source of typological asymmetries in frequency across pattern types. Another such question is whether observed properties of natural language need to be encoded as innately specified restrictions on linguistic structure, or whether these properties can be attributed to interactions between different cognitive biases, whether domain-general or domain-specific. Here, the source of the predicted typological biases demonstrated in this dissertation is attributed to differences in learnability across different patterns, and these differences in learnability emerge through the interactions between the grammar used to represent linguistic knowledge, and different learning dynamics observed in natural language acquisition.

In the case of the bias away from gang effect patterns, the learning agents' avoidance of cumulative constraint interactions would not be expected based on the structure of the grammar alone. For both the simple test case in Chapter 2 and the palatalization test case in Chapter 3, the predicted probability of the gang effect pattern in the baseline distribution exceeded at least one other pattern in the typology. The predicted probability of these gang effect patterns in the learning model results, however, was greatly reduced relative to other patterns in the typology, even though the learning algorithm had no specific incentive to avoid these patterns. In the case of the bias away from variation, the structure of the grammars used across the test cases in this dissertation did already contain a bias towards more deterministic grammars, as shown in the greater probability of more deterministic patterns in the baseline distributions. However, in the results of the learning models, the agents showed a tendency to avoid more variable patterns more extreme than would be expected based on the baseline distributions, even when tasked with learning variable patterns.

Explicitly encoding these biases into the structure of either the grammar or the learning algorithm, rather than modeling them as emergent biases, would result in a less explanatory model of these typological trends. Altering the structure of the grammar to disallow the representation of typologically infrequent patterns such as gang effects, or to disallow patterns of variation, would result in a grammar that undergenerates compared to the set of attested patterns. On the other hand, encoding a frequency or learnability parameter onto each pattern in a typology, in either the grammar or the learning algorithm, would still leave open the question of the source of this parameter and its value.

5.3.2 Conflicting and complementary biases

The iterated and interactive learning model results in this dissertation produce emergent learning biases that agree in direction, though these biases may differ in magnitude across models, with stronger effects tending to emerge in results of the interactive model. One question that may arise in the wake of this finding is, if both models produce similar effects on the predicted gradient typology, is it truly necessary to model both types of learning dynamics? Some previous work has shown examples where the iterated and interactive models produce different effects, suggesting that both dynamics are indeed necessary to produce a complete picture of the influences that shape language learning and language change (e.g. Dediu, 2009; Kirby et al., 2015; Smith et al., 2017).

Kirby et al. (2015), for example, show how semantic compositionality emerges from the combination of the learnability pressure exerted by the iterated model and the pressure for expressibility in communication exerted in the interactive model. In the extreme, the iterated model would favor one label for all possible meanings, which is easily learnable but not expressive, while the interactive model would favor a system with a unique label for each meaning, which is maximally expressive but more difficult to learn. In this case, the iterated and interactive models produce conflicting biases where the compromise is compositional structure, in which a small set of memorized labels can be recombined to assign a unique signifier for each meaning.

In contrast, for the phonological typology test cases presented in this dissertation, the iterated and interactive models do not produce conflicting biases, but rather complementary ones. Both models produce a bias away from gang effect patterns, and away from variation in the grammar. However, these are only two examples of typological tendencies out of a great many observed trends, such as those observed across the frequencies of different stress patterns (see e.g. Staubs, 2014; Stanton, 2016), or those observed across the structures of vowel inventories (see e.g. de Boer, 2000). Future work may explore and delineate the circumstances under which the predictions of the iterated and interactive learning models are conflicting or complementary.

One possibility is that differences between these models only emerge for the broad-scope core design features shared by all human languages, such as semantic compositionality, or perhaps the systematic structure exhibited in phonological inventories (see e.g. de Boer, 2000; Pater, 2012), in which case the typologies tested in this dissertation were too limited in scope to necessarily observe differences between the models. Another possibility is that the models may make different predictions if the learning agents need to infer the input from which the other agent generated their output form, rather than being trained on input-output pairs as they were in this dissertation. This inference step may introduce additional complexities in learning and communication that may produce different effects between the iterated and interactive learning models.

5.3.3 A combined learning model

As discussed in §1.4, the two agent-based learning models implemented in this dissertation each model a different type of learning dynamic observed in language acquisition. The iterated learning model simulates the unidirectional transmission of a language across generations of users, from adult models to child learners, while the interactive learning model simulates bilateral interactions between language users who are adapting to and learning from each other. A full model of natural language acqusition would ideally incorporate both types of learning dynamic; however, there are many challenging questions that arise in both implementing a combined model and interpreting the results, and so this dissertation, and much other work with the iterated and interactive learning models, simplifies the learning problem by focusing on the predictions that each learning model makes, independently of the other (though see Kirby et al., 2015, for a notable counterexample).

Pursuing a greater understanding of the effects of the iterated and interactive learning models individually will inform any future research into implementing a combined learning model. For example, in order to understand the results of a combined model, it is necessary to understand the effects of each component individually, and how each learning dynamic contributes to the overall result. Many of the open questions and challenges that arise in implementing a combined model, however, may only be answerable by making the attempt. Some of the questions that face future work in this area include whether there should be separate training and interaction phases, or whether these happen concurrently; whether the agents' grammars should become fixed at a point of "adulthood", or whether they should remain mutable until the agent is phased out; and whether the learner agents should weight data from the "adult" agents differently than data from peer learner agents. Any choices made in the design of a combined learning model could vastly alter the resulting predictions, and it remains to be seen which choices will most accurately reflect empirical observations.

5.3.4 Extendability of methodology

Although this dissertation has focused on the gradient typological predictions that arise from the combination of Maximum Entropy grammar and the iterated and interactive learning models, the overall methodology can be implemented using any system of representation encoding linguistic knowledge and any compatible learning model. Stanton (2016), for example, generates gradient typological predictions with a combination of parallel Optimality Theory and a convergent implementation of the Gradual Learning Algorithm in OT (Magri, 2012), and Culbertson et al. (2012) uses Bayesian modeling to generate predictions. Because it is so broadly applicable, this methodology can be utilized not only to investigate the results that emerge from a particular combination, as has been done in this dissertation, but also to compare gradient typological predictions between different grammatical theories or different learning models. This methodology, then, adds another way to test and explore the explanatory power of different theories and models.

5.4 Conclusion

In this dissertation, I demonstrate a method for generating gradient typological predictions through the combination of a theory of grammar and a theory of learning. Using specifically Maximum Entropy grammar (Goldwater & Johnson, 2003), combined with either of two agent-based learning models – the iterated learning model (Kirby & Hurford, 2002; Kirby et al., 2014) or the interactive learning model (see e.g. Pater, 2012) – I show how this system produces learning biases away from gang effect patterns, which rely on cumulative constraint interaction, and away from patterns of variation. Both of these biases mirror trends observed in natural language typologies, and emerge through the interactions between agents in the learning models, without needing to rely on restrictions encoded innately in the grammar. These results provide support for the use of this methodology in investigating the typological predictions of linguistic theories of grammar and learning, as well as in addressing broader questions regarding the source of gradient typological trends, and whether certain properties of natural language must be innately specified, or might emerge through other means.

BIBLIOGRAPHY

An, Gary, Qi Mi, Joyeeta Dutta-Moscato & Yoram Vodovotz. 2009. Agent-based models in translational systems biology. *Wiley interdisciplinary reviews: Systems biology and medicine* 1(2). 159–171.

Anttila, Arto & Giorgio Magri. 2018. Does MaxEnt overgenerate? implicational universals in Maximum Entropy grammar. *Proceedings of the 2017 Annual Meeting on Phonology* .

Arthur, W. Brian. 2015. *Complexity and the economy.* Oxford University Press.

Bane, Max & Jason Riggle. 2008. Three correlates of the typological frequency of quantity-insensitive stress systems. *Proceedings of the Northeast Linguistic Society 39 (NELS 39)* .

Bankes, Steven. 2002. Agent-based modeling: a revolution? *Proceedings of the National Academy of Sciences of the United States of America* 99(Suppl. 3). 7199–7200.

Bateman, N. 2007. *A crosslinguistic investigation of palatalization*: UC San Diego dissertation.

Boersma, Paul. 1997. Functional Optimality Theory. *Proceedings of the Institute of Phonetic Sciences of University of Amsterdam* 21. 37–42.

Boersma, Paul & Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25. 217–270.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32. 45–86.

Boersma, Paul & Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy & Joe Pater (eds.), *Harmonic grammar and harmonic serialism*, London: Equinox Press.

Bonabeau, Eric. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America* 99(Suppl. 3). 7280–7287.

Buchanan, M. 2009. Meltdown modeling: Could agent-based computer models prevent another financial crisis? *Nature* 460(7256). 680–682.

Burkett, David & Thomas Griffiths. 2010. Iterated learning of multiple languages from multiple teachers. *Evolang* 8.

Carroll, Lucien. 2012. Contrasts and contexts in palatalization. Ms. University of California, San Diego.

Coetzee, Andries. 2004. *What it means to be a loser: Non-optimal candidates in optimality theory*: University of Massachusetts Amherst dissertation.

Coetzee, Andries & Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.), *The handbook of phonological theory (2nd ed.)*, 401–431. Blackwell.

Culbertson, Jennifer & Paul Smolensky. 2012. A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science* 36. 1468–1498.

Culbertson, Jennifer, Paul Smolensky & Geraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition* 122. 306–329.

Daland, Robert, Andrea Sims & Janet Pierrehumbert. 2007. Much ado about nothing: A social network model of Russian paradigmatic gaps. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* 936–943.

Dale, R. & G. Lupyan. 2012. Understanding the origins of morphological diversity: the lingusitic niche hypothesis. *Advances in Complex Systems* 15, 3 & 4, 1150017.

de Boer, Bart. 2000. Self-organization in vowel systems. *Journal of Phonetics* 28. 441–465.

de Boer, Bart & Willem Zuidema. 2010. Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior* 18(2). 141–154.

Dediu, Dan. 2009. Genetic biasing through cultural transmission: Do simple bayesian models of language evolution generalise? *Journal of Theoretical Biology* 259. 552–561.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/.

Epstein, Joshua & Robert Axtell. 1996. *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press.

Ferdinand, Vanessa, Simon Kirby & Kenny Smith. 2019. The cognitive roots of regularization in language. *Cognition* 184. 53–68.

Finley, S. & W. Badecker. 2009. Artificial language learning and feature-based generalization. *Journal of Memory and Language* 61. 423–437.

Gafos, A. I. 1999. *The articulatory basis of locality in phonology*. Garland Publishing.

Goldstein, L. & C. A. Fowler. 2003. Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. S. Meyer (eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*, 159–207. Mouton de Gruyter.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. *Proceedings of the Workshop on Variation within Optimality Theory* 113–122.

Griffiths, Thomas L., Brian R. Christian & Michael L. Kalish. 2008. Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science* 32. 68–107.

Griffiths, Thomas L. & Michael L. Kalish. 2007. Language evolution by iterated learning with bayesian agents. *Cognitive Science* 31. 441–480.

Guy, Gregory. 1994. The phonology of variation. In Katharine Beals et al. (eds.), *Cls 30: Papers from the 30th regional meeting of the chicago linguistic society. volume 2: The parasession on variation in linguistic theory*, 133–149. Chicago Linguistic Society.

Hare, Mary & Jeffrey Elman. 1995. Learning and morphological change. *Cognition* 56. 61–98.

Hudson Kam, Carla. 2015. The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language* 91(4). 906–937.

Hudson Kam, Carla L. & Elissa L. Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1(2). 151–195.

Hudson Kam, Carla L. & Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59(1). 30–66.

Jäger, Gerhard. 2007. Maximum Entropy models and stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson & A. Zaenan (eds.), *Architecture, rules, and preferences: a festschrift for joan bresnan*, 467–479. CSLI Publications.

Jakobson, Roman. 1962. *Selected writings I: Phonological studies*. Mouton.

Jarosz, Gaja. 2013. Learning with hidden structure in optimality theory and harmonic grammar: Beyond robust interpretive parsing. *Phonology* 30. 27–71.

Jarosz, Gaja. 2016a. Computational modeling of phonological learning. *Annual Review of Lingusitics* 5.

Jarosz, Gaja. 2016b. Investigating the efficiency of parsing strategies for the Gradual Learning Algorithm. In *Dimensions of stress*, Cambridge University Press.

Johnson, Mark. 2007. A gentle introduction to Maximum Entropy Models and their friends. In J. Grimshaw, J. Maling, C. Manning, J. Simpson & A. Zaenen (eds.), *Architectures, rules, and preferences: a festschrift for joan bresnan*, 467–479. CSLI Publications.

Kelly, Michael. 1988. Phonological biases in grammatical category shifts. *Journal of Memory and Language* 27(4). 343–358.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure - an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5(2). 102–110.

Kirby, Simon, Tom Griffiths & Kenny Smith. 2014. Iterated learning and the evolution of language. *Current Opinion in Neurobiology* 28C. 108–114.

Kirby, Simon & James R. Hurford. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (eds.), *Simulating the evolution of language*, 121–147. Springer.

Kirby, Simon, Monica Tamariz, Hannah Cornish & Kenny Smith. 2015. Compression and communication in the cultural evoultion of lingusitic structure. *Cognition* 141. 87–102.

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1. 85–97.

Labov, William. 1994. *Principles of linguistic change, vol. 1: Internal factors*. Blackwell.

Labov, William. 2001. *Principles of linguistic change, vol. 2: Social factors*. Blackwell.

Labov, William. 2010. *Prinicples of linguistic change, vol. 3: Cognitive and cultural factors*. Wiley-Blackwell.

Legendre, Géraldine, Antonella Sorace & Paul Smolensky. 2006. The Optimality Theory - Harmonic Grammar connection. In Paul Smolensky & Géraldine Legendre (eds.), *The harmonic mind*, vol. 2, chap. 20, 339–402. MIT Press.

Lindblom, B., R. Diehl, S.-H. Park & G. Salvi. 2011. Sound systems are shaped by their users: The recombination of phonetic substance. In N. Clements & R. Ridouane (eds.), *Where do phonological features come from? cognitive physical and developmental bases of distinctive speech categories*, John Benjamins.

Lupyan, G. & R. Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5. e8559.

MacWhorter, J. 2002. What happened to English? *Diachronica* 19. 217–272.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29. 213–269.

Mansury, Y., M. Diggory & T.S. Deisboeck. 2006. Evolutionary game theory in an agent-based brain tumor model: exploring the 'genotype-phenotype' link. *Journal of Theoretical Biology* 238(1). 146–156.

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

Moore-Cantwell, Claire. 2015. The phonological grammar is probabilistic: New evidence pitting abstract representation against analogy. www.clairemoorecantwell.org/analogy/CMC_analogMS.pdf.

Moore-Cantwell, Claire & Joe Pater. 2016. Gradient exceptionality in Maximum Entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15. 53–66.

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25(1). 83–127.

Moreton, Elliott & Joe Pater. 2012a. Structure and substance in artificial-phonology learning, Part I: Structure. *Language and Linguistics Compass* 6(11). 686–701.

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

Pater, Joe. 2000. Non-uniformity in english secondary stress: the role of ranked and lexically specific constraints. *Phonology* 17. 237–274.

Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39. 334–345.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33. 999–1035.

Pater, Joe. 2010. Morpheme-specific phonology: constraint indexation and inconsistency resolution. In Steve Parker (ed.), *Phonological argumentation: Essays on evidence and motivation*, Equinox.

Pater, Joe. 2012. Emergent systemic simplicity (and complexity). *McGill Working Papers in Linguistics* 22(1).

Pater, Joe. 2016. Universal grammar with weighted constraints. In John McCarthy & Joe Pater (eds.), *Harmonic grammar and harmonic serialism*, Equinox Press.

Pawley, A. 2006. On the size of the lexicon in preliterate language communities: Comparing dictionaries of Australian, Austronesian, and Papuan languages. In J. Genzor & M. Buckov (eds.), *Favete linguis: Studies in honour of viktor krupa*, Institute of Oriental Studies.

Perfors, Amy & Daniel Navarro. 2014. Language evolution can be shaped by the structure of the world. *Cognitive Science* 1–19.

Peterson, G. E. & H. L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24(2). 175–184.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker. 2010. Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* 27(1). 77–117.

Prince, Alan. 2003. Anything goes. In Takeru Honma, Masao Okazaki, Toshiyuki Tabata & Shin ichi Tanaka (eds.), *New century of phonology and phonological theory*, 66–90. Kaitakusha.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar.* Blackwell.

Prince, Alan & Bruce Tesar. 1999. Learning phonotactic distributions. .

Pulleyblank, Douglas & William J. Turkel. 1996. Optimality Theory and learning algorithms: The representation of recurrent featural asymmetries. *Current trends in phonology: Models and methods* 2.

Rafferty, Anna N., Thomas L. Griffiths & Marc Ettlinger. 2013. Greater learnability is not sufficient to produce cultural universals. *Cognition* 129. 70–87.

Reali, Florencia, Nick Chater & Morten Christiansen. 2014. The paradox of linguistic complexity and community size. In E.A. Cartmill, S. Roberts, H. Lyn & H. Cornish (eds.), *The evolution of language: Proceedings of the 10th international conference*, World Scientific.

Reali, Florencia & Thomas Griffiths. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition* 111. 317–328.

Riggle, Jason. 2010. Sampling rankings. .

Rumelhart, D.E. & J.L. McClelland. 1986. On learning the past tenses of english verbs. In D.E. Rumelhart & J.L. McClelland (eds.), *Parallel distributed processing: Vol 2: Psychological and biological models*, MIT Press.

Samara, Anna, Kenny Smith, Helen Brown & Elizabeth Wonnacott. 2017. Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology* 94. 85–114.

Smith, Kenny. 2009. Iterated learning in populations of Bayesian agents. *Proceedings of the Annual Meeting of the Cognitive Science Society* 31.

Smith, Kenny, Amy Perfors, Olga Feher, Anna Samara, Kate Swoboda & Elizabeth Wonnacott. 2017. Language learning, language use, and the evolution of linguistic variation. *Phil. Trans. R. Soc. B* 372.

Smith, Kenny & Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. *Cognition* 116. 444–449.

Sonderegger, Morgan & Partha Niyogi. 2013. Variation and change in English noun/verb pair stress: Data and dynamical systems models. In Alan Yu (ed.), *Origins of sound change: Approaches to phonologization*, 262–284. Oxford University Press.

Stanton, Juliet. 2016. Learnability shapes typology: The case of the midpoint pathology. *Language* 92. 753–791.

Staubs, Robert. 2014. *Computational modeling of learning biases in stress typology*: University of Massachusetts, Amherst dissertation.

Stevens, K. N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17. 3–45.

Tesar, Bruce. 1998. An iterative strategy for language learning. *Lingua* 104. 131–145.

Tesar, Bruce & Paul Smolensky. 2000. *Learnability in optimality theory*. MIT Press.

Tesfatsion, Leigh. 2002. Agent-based computational economics: Growing economies from the bottom up. *Artificial Life* 8(1). 55–82.

Trudgill, P. 2011. *Sociolinguistic typology: Social determinants of linguistic structure and complexity*. Oxford University Press.

Walker, D.C., G. Hill, S.M. Wood, R.H. Smallwood & J. Southgate. 2004. Agent-based computational modeling of wounded epithelial cell monolayers. *IEEE transactions on nanobioscience* 3(3). 153–163.

Wedel, Andrew. 2011. Self-organization in phonology. In Marc van Oostendorp, Colin Ewen & Karen Rice (eds.), *The blackwell companion to phonology*, 130–147. Blackwell.

Wonnacott, Elizabeth. 2011. Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language* 65. 1–14.

Wonnacott, Elizabeth & Elissa Newport. 2005. Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M.R. Clark-Cotton & S. Ha (eds.), *Proceedings of the 29th annual boston university conference on language development*, Cascadilla Press.

Wonnacott, Elizabeth, Elissa Newport & Michael Tanenhaus. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology* 56. 165–209.

Wray, A. & G.W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117. 543–578.

Zipf, George K. 1949. *Human behavior and the principle of least effort*. Addison Wesley Press.

Zuidema, Willem & Bart de Boer. 2009. The evolution of combinatorial phonology. *Journal of Phonetics* 37. 125–144.

Zuraw, Kie. 2003. Probability in language change. In Rens Bod, Jennifer Hay & Stephanie Jannedy (eds.), *Probabilistic linguistics*, 139–176. MIT Press.