Computational Modeling of Learning Biases in Stress Typology

COMPUTATIONAL MODELING OF LEARNING BIASES IN STRESS TYPOLOGY

A Dissertation Presented

by

ROBERT DOUGLAS STAUBS

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 2014

Linguistics

COMPUTATIONAL MODELING OF LEARNING BIASES IN STRESS TYPOLOGY

A Dissertation Presented

by

ROBERT DOUGLAS STAUBS

Approved as to style and content by:

Joe Pater, Chair

John J. McCarthy, Member

John Kingston, Member

Sridhar Mahadevan, Member

John Kingston, Department Chair
Linguistics

For Helen and Bob

ACKNOWLEDGMENTS

Poised at the end of an academic journey stretching back decades, it is impossible to adequately state my gratitude and to acknowledge all those who have played a role. Despite this, I will press on and hope that the skeletal image I can construct gives some shape to my feelings.

Keeping with custom, I will begin with my committee. Joe Pater took me in from the start, bringing me into projects even in the summer before my first year. Like nothing else, his collaborative zeal made me feel like maybe I had something to contribute. Our meetings served to give all kinds of revelations, from how to properly attack a problem, to cutting through some disastrous prose. Things were always at least a little better after we met. This dissertation owes much and more to his inspiration, generosity with his time, and patient instruction. I appreciated all the times one of our arguments would polarize the other, bringing us both to better clarity in a parallel view of language. I am very grateful to have had Joe as my teacher and mentor and hope that I grow to better model myself on his example.

I feel lucky to have had the instruction of John McCarthy. The protean phonologist, he helps clarify my current ideas with one hand while sparking some new interest with the other. His encyclopedic knowledge of the field and phonological patterns is awesome to behold. I have never quite been sure whether it makes things more or less impressive when you realize he knows all these phenomena because he has written on them. John has a way of understanding your argument before you have even made it, and sometimes before you have even thought it—very useful for getting through rough patches. John has always been a critic of my writing, but he manages to do so without seeming like a critic of the thought it represents. For this,

and for the (incremental) progress I hope I have made, I thank him. John has been available numerous times to push me on professional advice even as I ask the most naive things. It has been an honor and a privilege.

John Kingston has been very generous with his time even as I come to him with ideas half-formed. He has tolerated all manners of phonetic naivete on my part, helping me see the kernel of possible insight hiding beneath the bewilderment. I cannot say that I have overcome all these issues as yet, but John has been exceedingly helpful in knowing where to look. He brings a piercing gaze on the obfuscations and oversimplifications that can plague phonetic and phonological argumentation. If I should learn even a fraction of that, I should be content. I thank him immensely for his time and energies.

I am grateful to members of the McCarthy/Pater Grant Group, Phonology Reading Group, and Sound Seminar at UMass for comments and feedback on this work. I also owe a lot to people outside UMass. Thanks in particular to Adam Albright, Eric Baković, Ryan Bennett, Jeff Heinz, René Kager, and Anne-Michelle Tessier for insights at various stages of development of the work presented in this dissertation. Max Bane kindly provided me with the parsed database information that started me off on some of the work in Chapter 2. Thanks also to audiences at NECPhon IV, the 2nd UConn Workshop on Stress and Accent, WCCFL 31, RUMMIT IV, and mfm 22 for helpful comments.

Thanks to the National Science Foundation, from which I received support through much of this work. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0907995. I also received financial support during parts of this research from Grant BCS-0813829 from the National Science Foundation to the University of Massachusetts Amherst.

I owe many various types of thanks to all the other UMass faculty who taught me or otherwise helped me in my time at UMass, whether for courses taught, generals

papers advised, or any other reason. A partial list of the people I was blessed with includes: Emery Berger, Rajesh Bhatt, Seth Cable, Andrew Cohen, Brian Dillon, Lyn Frazier, Lisa Green, Alice Harris, Neil Immerman, Kyle Johnson, Erik Learned-Miller, Sridhar Mahadevan, Barbara Partee, Anne Pycha, David Smith, Peggy Speas, Ellen Woolford, and Kristine Yu. Thanks also to Kathy Adamczyk, Tom Maxfield, Michelle McBride, and Sarah Vega-Liros for help with all the administrative crises I managed to invent.

I would not have ended up at UMass if not for the support I had in my undergraduate degree at the College of William and Mary. Among the many: Anne Charity Hudley, for bringing me into her lab; Jack Martin, for advising me; Ken Lacy, for giving me my unlikely introduction to Optimality Theory and modern theoretical phonology; Anya Lunden, for furthering that education and inspiration; and Ann Reed, for being Ann Reed. I give short shrift to these people and elide others (notably in my other home of Computer Science). I hope I can be forgiven this fault, as they have graciously forgiven others.

Moving to the slightly less formal academics, my cohort. Eight of us arranged ourselves into 311 South College our first year, and most of us were still in that room when Linguistics moved out of South College, just this summer. Our starting roster: Elizabeth Bogal-Allbritten, Minta Elsman, Seda Kan, Claire Moore-Cantwell, Presley Pizzo, Jason Overfelt, and Andrew Weir. Without them, I would have had a much harder time adjusting to graduate school, and would have had much less fun. My particular thanks to Andrew, Claire, and Minta, for living with me first year at 40 Grant and helping convince me I had made a good decision. Thanks to all the other grad students and affiliates who made this time special: Chris Davis, Lena Fainleib, Meg Grant, Hannah Greene, Clint Hartzell, Ivy Hauser, Matt Hine, Coral Hughto, Karen Jesney, Mike Key, Wendell Kimper, Nick LaCara, Josh Levy, Kevin Mullin, Alex Nazarov, Kat Pruitt, Amanda Rysling, Anisa Schardl, Shayne Sloggett, Brian

Smith, Megan Somerday, Guillermo Vales Kennedy, Martin Walkow, and all others that I have inevitably neglected to mention. Work parties with Anisa, Claire, and others in these lists helped me through a lot.

At points in these five years, it has been useful to have a place to step back to in my larger life. Having generally failed at creating a place free of linguistics in Massachusetts, it has been very valuable to have friends in the wider world—here principally defined as "Northern Virginia." Thanks to Tony, my friend since age thirteen despite only having one class together. Thanks to the friends I made in high school for leaving me a place to fit even as I was gone longer and longer: Curtis, Garrett, Kurt, both Matts, and more. Thanks to Mike and Chad for our continued rapport developed as undergraduates. Thanks to Chad particularly for responding "sure" when I jokingly suggested he move up from Florida to live with me.

I cannot adequately thank Barbara and David Staubs, my mother and father, for their love and support in my academic life and beyond. The completion of this dissertation is in no small part due to their support. Without the freedoms, persistence, and value of learning they instilled in me from an early age, I do not think I would be in the same place. Thank you, mom and dad.

My grandparents, Helen and Robert Spence, have a foundational importance to my academic life. I feel I did not appreciate this adequately early on, but I think the signs are clear now. They were both academics: Helen in computer science and Bob in physics. Though his vocation would not show it, Bob was always interested to learn of a new etymology or some quirk of language. Ending up with a dissertation in computational phonology after early exposure to academics interested in computation and language seems too much of a coincidence to ignore. I dedicate this dissertation to them.

I met Alexandra Neyers very near to the start of graduate school. As such, my journey has been very much a shared one. Her support has helped push me through

the many steps along the way that could have been true roadblocks. Her strength and compassion has been a continual inspiration. Ali, thank you for walking alongside me on this path.

I have had an amazing set of experiences and met many wonderful people in the process of writing this dissertation. Actually getting it done? Well, that's lagniappe.

ABSTRACT

COMPUTATIONAL MODELING OF LEARNING BIASES IN STRESS TYPOLOGY

SEPTEMBER 2014

ROBERT DOUGLAS STAUBS

B.S., THE COLLEGE OF WILLIAM & MARY IN VIRGINIA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Joe Pater

This dissertation demonstrates a strong connection between the frequency of stress patterns and their relative learnability under a wide class of learning algorithms. These frequency results follow from hypotheses about the learner's available representations and the distribution of input data. Such hypotheses are combined with a model of learning to derive distinctions between classes of stress patterns, addressing frequency biases not modeled by traditional generative theory.

I present a series of results for error-driven learners of constraint-based grammars. These results are shown both for single learners and learners in an iterated learning model. First, I show that with general n-gram constraints, learners show biases in their learning of stress patterns, mirroring frequency effects in the observed typology. These include biases toward full alternation and fixed stress near word edges. I show that these effects arise from the learner's representation of the consistency and

distinctiveness of learning data. I formalize this notion within error-driven, constraint-based learners.

I show how specific representational assumptions can lead to distinct predictions about frequency, potentially adjudicating between theories. Languages with primary stress placement independent of word parity are shown to be—with the right constraint set—more consistent and thus more readily learned, offering an explanation for their relative frequency. This explanation is especially valuable because, while parity-dependent languages exist, they are a small minority. I continue by showing how such a model predicts biases in the size of stress windows and discuss the role of this approach in deciding the nature of potentially "accidental" gaps.

I demonstrate that such a model can incorporate sources of bias outside the learner's representations. I give a model of a perceptual nonfinality effect based on probabilistic misperception. This modification is shown to help account for typological skews in the edge of fixed stress and windows, as well as foot type for iterative stress.

The methods used and conclusions drawn in this dissertation are potentially extendable to a wide range of linguistic phenomenon. This foundation is a way of approaching some otherwise-unexplained frequency biases by grounding them in theories of linguistic representation and learning.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

FORMAL BIASES IN STRESS LEARNING

1.1 Overview

In this chapter, I first introduce the problem of probabilistic biases in linguistic typology. Not all typological generalizations are categorical in nature, distinguishing only between what is possible and what is impossible. Instead, many generalizations are *tendencies*: statements of which sorts of pattern are more or less common. These tendencies are problematic for generative phonology because the typical types of models used are ones which make only categorical predictions, distinguishing between systems the grammatical theory can represent and those it cannot. I argue for a useful division of labor between the grammatical theory and a learning theory: the grammatical theory provides a representational space in which learning operates, biasing learning towards or away from particular patterns. I briefly sketch the use of such biases to explain typology in an iterated learning model.

I next introduce the models to be used throughout: Maximum Entropy grammar (Goldwater and Johnson, 2003), Robust Interpretive Parsing (Tesar and Smolensky, 2000), and one interpretation of iterated learning.

I discuss models of two types of bias. The first concerns the biases that emerge from a theory of stress using very general (n-gram) constraints. Such a theory is shown to predict biases in the frequency of iteration in stress and its most typical form. I explain the emergence of this bias in terms of the key concept of *distinctiveness*. The second type of bias, probabilistic predictions dependent on assumptions of featural representations, serves to elaborate on these concepts.

Finally, I formalize distinctiveness and related concepts. I provide the mathematical background behind the biases discussed in the dissertation. I explain the connection between error-driven models of learning in Maximum Entropy (Jäger, 2007; Boersma and Pater, 2014) with classic work on perceptron learning (Novikoff, 1962). I prove conditions under which a pattern will be favored or disfavored by such a learner.

1.2 Introduction

The principal goal of grammatical theories, particularly in a generative framework, is to accurately predict the attested range of human linguistic systems. In its most typical form, this work consists of matching predictions to a categorical typology. Languages are described as attested or unattested, values of 1 or 0, and the grammatical theory is designed to divide the space of logically possible languages along these lines. Thus attested languages are meant to be representable in the hypothesis space of the theory and unattested languages should not be. This approach to typology and grammatical theorizing has driven much of the typological work in the generative tradition, especially with frameworks like Optimality Theory (OT; Prince and Smolensky, 1993/2004) and its Principles and Parameters predecessors (Chomsky, 1979).

I argue that a categorical approach to linguistic theory is not rich enough. The categorical view of typology ignores something crucial about linguistic patterns as they actually are observed: they have frequencies. An attested pattern need not be common—indeed, many patterns are apparent singletons. Standard generative theory allows only two choices in dealing with such extremely rare patterns. They can either be treated on a par with better attested patterns, or they can be ignored. In many cases, this choice is arbitrary, and neither one is fully satisfactory. A concrete example is given in §2.2. Languages in which main stress is placed on the "last" foot

in an iterative stress system are very rare, but theories typically treat these on a par with the usual placement of main stress on the "first" foot, and prior attempts to address the observed difference have not succeeded.

This probabilistic view of the data given in typology corresponds to a probabilistic view of grammatical theorizing. Rather than aiming for a theory that separates attested from unattested, we can aim for the richer goal of predicting relative frequencies. The typology of linguistic stress provides a useful domain in which to study modeling of probabilistic typology, as previously undertaken by Bane and Riggle (2008).

Stress is a common property of a variety of languages. It has been the subject of a number of extensive typological studies (e.g. Hyman, 1977; Heinz, 2007; Goedemans, 2010), allowing for a reasonable understanding of the frequency of both individual patterns and overall trends. In stress there are obviously common patterns, such as penultimate stress, and obviously less common ones, such as antepenultimate stress. We also see a divide between types of gaps: some unattested languages seem plausible, such as four-syllable stress windows or pre-antepenultimate stress; others seem simply impossible as human languages, such as stress on every prime-numbered syllable. Simplifying these numerical patterns to a categorical distinction loses much of the information in a survey. Ideally, our linguistic theories should, *in toto*, explain as much of the observed pattern to human linguistic variation as possible. If we exclude frequencies, losing the information they contain, we are no longer even attempting this ideal. To better make such an attempt, our linguistic theories should be ones which make predictions about frequencies, tested against frequency data. Stress data allows us to verify these kinds of predictions with relative ease.

A theory of probabilistic typology need not take us far afield from the typical assumptions of analysis. Representational assumptions are always necessary for any theory of grammar and linguistic variation because they are necessary for the learn-

ing of structure. In the generative tradition, these assumptions are substantial and likely specific to the language faculty. These assumptions encode the task-specific prior knowledge concerning the possible range of linguistic variation. Representations do not logically need to be so specific, but without *some* initial assumptions, a learner can never make progress toward an effective hypothesis about the language to which it has been exposed. In addition to representation, there is another set of assumptions that are typically omitted in linguistics analysis: assumptions on the mechanics of learning. However, again, language *is* learned and thus these assumptions are required. With just these two components—representation and learning—we can extract probabilistic predictions. Any non-trivial learning algorithm, paired with a representation, will exhibit biases for or against particular patterns described within that representational space. Thus, at least some probabilistic predictions emerge automatically from components that are independently needed (and used) for a theory of linguistic structure. Given that such biases must exist, it is natural to start with learning in developing a theory of non-categorical typologies.

Stress typology proves a fruitful place to focus for this particular tack. Stress generalizations operate over comparatively abstract phonological elements and descriptions such as "stressed," "unstressed," "heavy," and edge alignment. With a few exceptions (e.g. interactions of weight and sonority with stress, nonfinality, etc.), the descriptions of stress are relatively divorced from the phonetic substance. Although surely perception and production have an effect on stress typology, their effect is felt much less profoundly than in many other domains of phonological structure. Thus stress provides a useful testbed for modeling probabilistic patterns using learning: it is not seemingly possible, for example, to attribute the typology of stress to perception and production *per se*. The substantial number of tendencies which remain unexplained in a standard generative approach demand explication: as developed here, through learning.

1.3 Grammar, learning, and typology

1.3.1 Grammatical assumptions

In order to establish distinctions between linguistic patterns in their relative learnability, it is useful to be able to inspect the results of learning and assign more than one value to the results—not just a success or failure. One way in which this can be accomplished is by using a stochastic grammar model. In such a system, the goal of learning is to match the probability distribution of the input data, which might happen to be categorical in nature. These grammars induce a probability distribution over a set of possible output forms given a particular input. The process of learning a *categorical* grammar consists of giving the target forms probabilities closer and closer to 1.

I adopt Maximum Entropy Grammar (MaxEnt; Goldwater and Johnson, 2003) as a formalization which satisfies this criterion. MaxEnt is a form of Harmonic Grammar (HG; Legendre et al., 1990; Smolensky and Legendre, 2006), establishing probabilities of output forms on the basis of their weighted sum of violations. The grammar maps a set of inputs to candidate outputs through the operations Gen, and these output candidates are assigned violations based on a constraint set Con. MaxEnt thus shares much of the framework of Optimality Theory, only differing importantly in its assignment of probabilities and use of weights. The merits of probabilities are clear for the kind of work undertaken here—probabilistic predictions are ideal for probabilistic data. The weighted constraints of Harmonic Grammar have been extensively explored, yielding potentially positive and negative aspects of such a model (e.g. Pater, 2009). My use of MaxEnt suffices to describe probability distributions simply within a framework that is relatable to the assumptions discussed in much of generative phonology.

In MaxEnt, output probabilities are taken to be proportional to the exponential of harmony, itself the weighted sum of violations for a candidate.

$$p(j|i) \propto e^{H_{ij}} \tag{1.1}$$

$$= e^{\mathbf{w}^T \mathbf{v}_{ij}} \tag{1.2}$$

H_{ij} refers to the harmony of candidate j coming from input i with violation vector v_{ij} under the weights w. A violation vector is an ordered set of numbers corresponding to the violations a candidate incurs for each constraint in the constraint set Con.

The coefficient of proportionality Z_i just serves to create a probability distribution out of this quantity. It is the sum of exponentiated harmonies for all candidates in the set Gen(i)—all candidates generated from the input y under the generator Gen.

$$p(j|i) = \frac{1}{Z_i} e^{H_{ij}} \tag{1.3}$$

$$Z_i = \sum_{y \in \text{Gen(i)}} e^{H_{iy}} \tag{1.4}$$

Violations are assumed to be non-positive and weights non-negative. Thus, worse candidates have more negative harmonies, leading to lower probability. For example, consider two candidates a and b with harmonies -1 and -4, respectively. The exponentials of these harmonies are $e^{-1} \approx 0.37$ and $e^{-4} \approx 0.02$. To calculate the probability of candidate a we normalize, dividing this value by the sum of both: $\frac{0.37}{0.37+0.02} \approx 0.95$.

MaxEnt is not unique in being a constraint-based stochastic model of grammar. For example, Stochastic Optimality Theory (Boersma, 1997) induces probability distributions over candidates by using numerical ranking values and random noise. Noisy Harmonic Grammar (Boersma and Pater, 2014) is similar—harmonies are evaluated as normal in Harmonic Grammar except that the constraint weights used are perturbed by random noise on each evaluation. In this study I have chosen MaxEnt

over e.g. Noisy HG for its simple and explicit method of calculating candidate probabilities. MaxEnt, in contrast to alternatives, does not require sampling in order to form a probability distribution and is grounded in a large background in statistics and machine learning (e.g. Berger et al., 1996).

1.3.2 Learning

1.3.2.1 Update rules

To obtain a relative view of learnability for various languages it is necessary to have an explicit model for learning. The learning model used here is online—the learner processes each datum it receives in turn, adapting its hypothesis. A "teacher" randomly selects an input and produces an output based on its grammar. The learner considers a candidate set consisting of all candidates that are generated by the input. The learner produces its own output for the input. If the learner's predicted output does not match the teacher's, the learner updates its constraint weights. Learning is therefore error-driven: updates occur when expected and observed data do not match. The rule could be modified to perform an update even on success, but as shown below, updates yield no change to the weights given a match between the learner's production and the teacher's.

Stated another way, the teacher samples inputs i from a distribution $\mathbf{d}$ and produces outputs j^* such that the pair (i, j^*) is in the teacher's grammar. This pair is given to the learner, which uses its current grammar $\mathbf{w}^t$ to produce an output j for i. If j and j^* differ, $\mathbf{w}$ is updated.

The learner's grammar is updated according to Stochastic Gradient Ascent (SGA; Jäger, 2007). This update is essentially the same as Harmonic Grammar Gradual Learning Algorithm (Boersma and Pater, 2014), and known elsewhere in slight variations variously as the perceptron update rule (Rosenblatt, 1957) or delta rule (Widrow and Hoff, 1960). In this update rule, the old weights are adjusted by the difference

Overt form	Consistent hidden structure
$\sigma\acute{\sigma}\sigma$	$(\sigma\acute{\sigma})\sigma$ $\sigma(\acute{\sigma}\sigma)$
$\sigma\acute{\sigma}\sigma\acute{\sigma}\sigma$	$(\sigma\acute{\sigma})(\sigma\acute{\sigma})\sigma$ $\sigma(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)$

Table 1.1: Ambiguity in hidden foot structure between iambic and trochaic parses.

between the violations of the learner's chosen candidate and the violations of the teacher's chosen candidate, scaled by a learning rate η.

$$\text{New Weights} = \text{Old Weights} + \eta \times (\text{Teacher Violations} - \text{Learner Violations}) \quad (1.5)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} + \eta(\mathbf{v}_i^* - \mathbf{v}_{ij}) \quad (1.6)$$

This update increases the weight of constraints violated more in the learner's erroneous form, penalizing such violations more heavily. It decreases the weight on constraints violated more in the teacher's chosen form, permitting such violations more. If there is no difference in violations for a particular constraint, no change is made. This update could potentially produce negative weights, creating a "benefit" constraint rather than a penalty. This potentially subverts the typological motivations of such a constraint set (see e.g. Pater, 2009) and is thus not permitted in the simulations adopted in this work. If a constraint weight would become negative by this update, it is instead set to zero. Thus the best a violation can ever do for a candidate is to not affect its probability one way or another—violations can never help.

Not all types of phonological structure are immediately known to a learner on the basis of a surface form. Hidden structure problems arise when there is ambiguity in the interpretation of the full structure of an overt form. For example, the foot structure of a three-syllable word with medial stress is ambiguous. The hidden structure could be

(at least) a left-aligned, right-headed foot or a right-aligned, left-headed foot. This is shown in Table 1.1. For some of the simulations presented, hidden structure is a concern. This is particularly the case for stress simulations featuring constraints which make reference to foot structure. Without a foot structure, violations cannot be assessed and the learner cannot compare its own output with the teacher's. The learner must therefore make some decision about what hidden structure to use in evaluating the teacher's constraint violations. The approach presented here uses a probabilistic adaptation of Robust Interpretive Parsing (RIP; Tesar and Smolensky, 2000; Boersma, 2003; Jarosz, 2013; Boersma and Pater, 2014) to choose a likely hidden structure. In this version of RIP, the hidden structure used for a particular overt form is probabilistically chosen according to the grammar from all hidden structures consistent with the form. Thus the learner picks a foot structure compatible with the teacher's form with a probability related to the learner's own assessment of the well-formedness of full structures. This approach bears an essential similarity to the general Expectation-Maximization approach to hidden structure (Dempster et al., 1977).

$$p(k|i,j) = \frac{1}{Z_{ij}} p(j|i,k) \tag{1.7}$$

$$Z_{ij} = \sum_{q \in \textsc{Interp}(\mathrm{i,j})} p(j|i,q) \tag{1.8}$$

$\textsc{Interp}(i,j)$ gives the set of full hidden structure interpretations of the overt output j with input i—that is, the set of hidden structures logically consistent with an observed overt structure.

An important concern for the convergence of SGA is the size of the learning rate. If the learning rate is very small, the learner cannot move its hypothesis far for any given prediction error. This means that any convergence will take longer than might

be necessary. In contrast, if the learning rate is too high, the learner might "skip over" the point at which it would converge. In this case it would, for example, push a constraint weight from being too low to too high, rather than simply decreasing distance to the ideal value. Because of these kinds of concerns, it is important to test more than one learning rate for a learning problem. In this dissertation, I present results using single learning rates. However, for each set of data presented at least one additional learning rate was tested. Most typically, a learning rate an order of magnitude higher and one an order of magnitude lower were tested. Thus if the text mentions a learning rate of $\eta = 0.1$, typically learning rates $\eta = 1.0$ and $\eta = 0.01$ were also tested. For the types of conclusions discussed in this work, no important differences were found for learning rate. While learning rate can make convergence faster or slower, it does not do so in a way that biases for or against particular languages in the cases discussed.

1.3.2.2 Sampling

The random sampling of input-output pairs from the teacher represents the data available to the learner from its environment. This distribution $\mathbf{d}$ is of potential interest because the frequency of data could have an effect on the relative learnability of patterns. For example, the learnability of a stress language could be affected by the length of words needed to disambiguate it. In an antepenultimate stress system, stress falls two syllables from the right edge of the word. In one-, two-, and three-syllable forms, two syllables from the right edge is the *left* edge. Therefore, in any word under four syllables, antepenultimate stress is ambiguous with initial stress. We should expect that this ambiguity poses a problem for the learner. The distribution over word lengths is therefore quite important: if long words are uncommon, the ambiguity is pervasive. If, in contrast, long words are relatively frequent, it is not as

much of a concern. For this sort of reason, the sampling distribution over inputs is of interest in the experiments presented.

In the stress cases discussed, the sampling distribution is some form of exponential distribution relating the number of syllables to the probability. This mirrors the distribution in, for example, child-directed speech in English, as seen in CHILDES (MacWhinney, 2000). Figure 1.1 shows the results of computing a distribution of word length counts based on an onset maximization algorithm *à la* Dell and Elmedlaoui (1985) using pronunciations from CMUdict (Weide, 1994) for words with CHILDES counts. Thus short words are sampled exponentially more often than words of longer lengths. This is to avoid the sharp mismatch with reality that would result if the learner were consistently provided with very long words (as would be the case with, for example, a uniform distribution over some range). In these cases the sampling distribution is also artificially limited to include only up to a certain maximum word length, ranging in particular between one and eight syllables (a very slight modification to the numbers already present in the CHILDES data). This is more typically limited to a two to eight syllable range as monosyllables offer no evidence of errors under an assumption of culminativity, under which any string must contain at least one stress.

$$\mathbf{d}_n = \frac{e^{-kn}}{\sum_{m=1}^{8} e^{-km}} \tag{1.9}$$

I do not take the English values as exactly representing the cross-linguistic distribution across word lengths. This must surely vary for a variety of reasons, including varying morphological structure and minimal word restrictions. Instead, the English data points to a general functional form for this word length distribution which should generally approximate the kind of differences found across word lengths in a variety

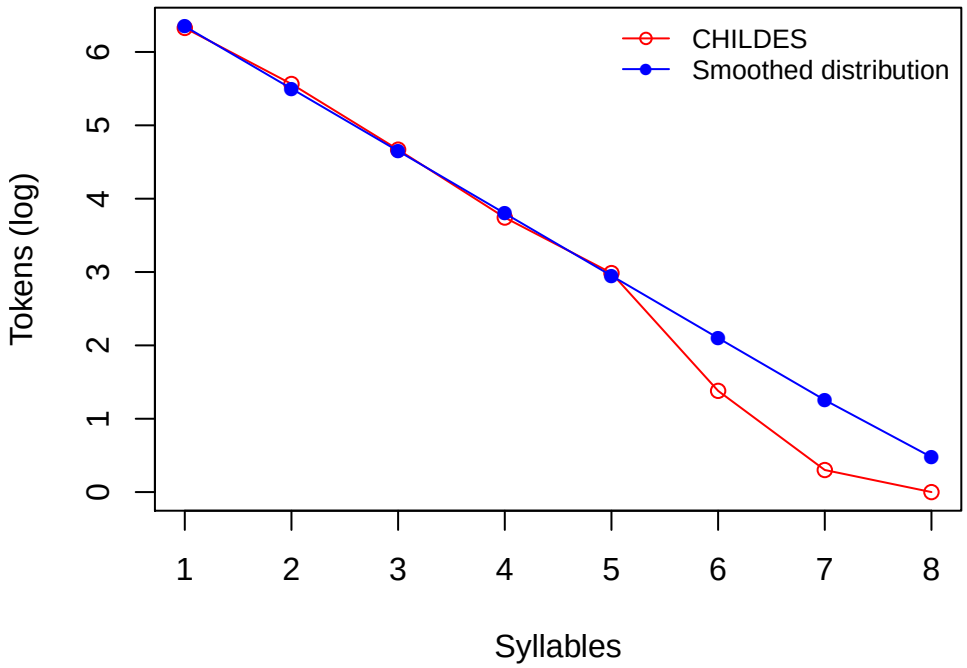


Figure 1.1: Sampling distribution for learning simulations produced by fitting an exponential curve to word length data derived from CHILDES.

of languages. Additionally, the biases presented in this work are largely robust to distribution—additional tests with the uniform distribution (for example) provide similar qualitative results.

1.3.3 Iterated learning

Learning results alone do not immediately inform us about typology. A single instance of transmission from a teacher to a learner cannot hope to realistically shape the whole of human language in the ways we might be concerned with when examining frequency. For that to be the case, a learner would need to be able to routinely depart widely from the language of its teacher. One place in which we do see such departures is in reanalysis (see e.g. Harris and Campbell, 1995, on cases of syntactic reanalysis). However, these cases crucially involve a reinterpretation of covert structure, such that a difference between the teacher's grammar and the learner's is not readily apparent. If, as seems reasonable, learners must largely agree with their teachers on the form

of their language, learning biases must not radically shift output strings and must instead have a more subdued effect on typology.

One way in which this effect could be made manifest is through iterated learning (e.g. Kirby, 2002; Griffiths and Kalish, 2005), reviewed also by (e.g. Zuraw, 2003; Wedel, 2011) and contrasting with paradigms such as social learning (Niyogi and Berwick, 2009). In an iterated learning configuration, a learner acquires its language from a teacher (or, more broadly, teachers) and then must serve as the teacher for other learners (Figure 1.2). If we then compare the distribution of languages in the first generation of learners with one in the future, we see a potential for small learning biases to be amplified. Each generation of learners can make a small change, altering only a little of its language while maintaining gross agreement with its teacher. Added up over many generations, however, this progression can yield large deviations directed by those individual learning biases. Work by Kirby and colleagues is largely focused on the emergence of universals, but probabilistic typology is not far removed from this view.

A body of recent research by Griffiths and colleagues has focused on the iterated learning of language and other culturally transmitted concepts. This work is in a Bayesian perspective. Each learner responds to data according to Bayes' theorem, updating the probabilities given to data (and therefore, to languages), based on a combination of its input data and its prior beliefs about the probability of languages. If learners form a single chain, with one learner teaching another, results are as analyzed by Griffiths and Kalish (2007). Two options are considered: learners which select their language by sampling from the posterior distribution derived from Bayes' theorem ("samplers") and learners which select the maximum of the posterior distribution (*maximum a posteriori*, MAP).

A chain of samplers is shown to be equivalent to a Gibbs sampler. The posterior distribution of such a chain therefore converges to the prior distribution. That is, a

Learner 1 → Learner 2 → Learner 3 → Learner 4 → ...

Figure 1.2: Basic setup for iterated learning.

uniform population of such chains, past convergence, would show statistics directly reflecting the biases of the prior distribution of individual learners. MAP learners also converge towards the prior, but do not mirror it. This learning configuration is equivalent to Expectation-Maximization with the learner's data serving as latent data and no observed data modeled. The posterior distribution of such a chain will have a maximum at the maximum of the prior, but other aspects of its shape are controlled by the speed of changes between generations, influenced by transmission factors and properties of the individual languages. In this case, a uniform population of such chains will resemble the prior in its most preferred languages, but not in all respects. Kirby et al. (2007) show an infinite continuum of agent behaviors between samplers and maximizers.

Griffiths and Kalish (2007) also show that a generalization of these kinds of models yields population convergence to the prior. In this model, there are an infinite number of learners acquiring their language from randomly selected teachers. In this paradigm, the fraction of learners holding to any one language converges to the prior probability of that language. Dediu (2009) discusses another alternative paradigm in which generations consist of heterogenous pairs of learners. In this setting samplers are less distinct from MAP, both largely converging to the prior.

All of this goes to show that the model of iterated learning *matters*, and that it might not directly mirror the biases in a prior. However, we also see biases of one model of iterated learning echoed in another, with numeric differences. These sorts of qualitative comparisons are possible as overviews, and direct the kinds of explorations developed in this work.

In this dissertation, I consider only the simplest model of iterated learning, in which a single learner acquires its language from a single teacher. Typological statistics are therefore computed over many trials and potentially many different starting positions (i.e. initial languages or grammars of the first teacher). Within this approach I take two tacks. The first is a direct implementation of iterated learning, in which the final state of the learner after some amount of learning serves as the exact starting state of a teacher in the next generation. In the second approach, the transition between one generation and the next is taken as a probabilistic change between categorical language states. The statistics of one generation of learning are computed and used to derive theoretical outcomes over many generations. This can be done quite simply by construing the transitions between one generation and the next as a stochastic transition matrix and exponentiating this matrix for the number of generations required.

As an example, consider Table 1.2. The first matrix shows the results after an imagined measurement of a single generation of learning (e.g. simulation). There are three languages, with each row representing the distribution of a learner's ultimate language when its teacher had a given language. Each row will sum to one because the learner must learn one of the three languages. Language 1 is typically learned faithfully, Language 2 is learned as some language at chance, and Language 3 is learned most typically as some language other than Language 1. What happens when the learner acts as learner to a new generation? The second matrix informs us that all starting positions are more likely to end up with a hypothesis of Language 1 than before—this makes sense, because Language 1 was the language most faithfully learned. Finally, the last matrix shows us the long-term behavior of this system. Language 1 tends to dominate, with Languages 2 and 3 equally common. This is the most natural view of the "typological" predictions of the initial learning result matrix.

$\mathbf{M}^1$		Result after 1 generation		
		Language 1	Language 2	Language 3
Starting language	Language 1	0.900	0.050	0.050
	Language 2	0.333	0.333	0.333
	Language 3	0.100	0.450	0.450

$\mathbf{M}^2$		Result after 2 generations		
		Language 1	Language 2	Language 3
Starting language	Language 1	0.832	0.084	0.084
	Language 2	0.444	0.278	0.278
	Language 3	0.285	0.358	0.358

$\mathbf{M}^{1000}$		Result after 1000 generations		
		Language 1	Language 2	Language 3
Starting language	Language 1	0.684	0.158	0.158
	Language 2	0.684	0.158	0.158
	Language 3	0.684	0.158	0.158

Table 1.2: Stochastic matrices representing hypothetical results of learning (1 generation) and projections for iterated learning (2 generations and 1,000).

My iterated learning model assumes three things. First, the languages of interest are a finite number of categorical states: any teacher has one and only one language it teaches to its learner, and this language is describable without the use of probabilistic grammar. Second, to reach this state learners must pick the maximum likelihood language corresponding to their probabilistic hypothesis. Finally, the population frequencies of given language types correspond to the probabilities of each of these states resulting from a single chain.

Importantly, I do not assume neither that the distribution of languages in the real world nor the distribution over languages for a chain of learners has converged. This caution is motivated by work by Rafferty et al. (2009) suggesting that ecological convergence is unlikely to have occurred in the world's languages. For numerical estimates I therefore numerically fit the number of iterations to observed data (Chapters 2 and 3), rather than assuming the stationary state of a Markov chain.

I return to these kinds of models in Chapters 2 and 3, but it is worth emphasizing always that a single learning result cannot predict typology on its own, as discussed by e.g. Rafferty et al. (2011). However, a learning model placed in such a view of language change *can* be enabled to make predictions.

1.3.4 Comparison with evaluation metrics

Chomsky and Halle (1968) advance an evaluation metric for phonological grammars (§§8.1, 8.A). The evaluation procedure is important in cases of ambiguity: if more than one grammar is compatible with the observed linguistic data, which grammar should the learner choose? Chomsky and Halle propose that the learner seeks to maximize an evaluation metric, defined as the reciprocal of the minimal size of the grammar. That is, learners faced with a choice that cannot be resolved on empirical grounds make the decision based on a concern for parsimony.

This procedure on its own serves to select "simple" grammars, but cannot make distinctions in terms of phonetic naturalness. To address this, Chomsky and Halle augment the metric in §9.2. In this instantiation, each feature is given a marked and an unmarked value. Unmarked values do not contribute to the size of a grammar, and therefore the most preferred grammars will be those with the lowest number of unmarked features. Modified in this way, the evaluation procedure prefers both simpler grammars and grammars in greater accord with observed patterns of phonetic naturalness.

Chomsky and Halle intend the evaluation procedure to be, at least in part, a theory of frequency. Grammars which are preferred by the evaluation metric are assumed to correspond roughly with more frequent linguistic patterns:

> We would expect, naturally, that systems which are simpler, in this sense, will be more generally found among the languages of the world, will be more likely to develop through historical change, etc. (Chomsky and Halle, 1968, p. 411)

An evaluation metric alone cannot produce frequency predictions. The metric itself is just a number scoring a grammar—it cannot by itself influence the frequency of the system. Simplicity alone is not predictive in generative phonology, despite occasional intuitions to the contrary. I discuss one such case in §1.6.3. The evaluation procedure, however, operationalizes the metric. The procedure accepts or rejects grammars on the basis of the metric's value, offering an opening for linguistic change if learners fail to receive exhaustive disambiguation of their languages (see Bach and Harms, 1972, on relations between the procedure and historical change).

In this view, the evaluation metric approach to typological frequency is not dissimilar from my own. The evaluation procedure induces systematic, biased mislearning given insufficient data. This is comparable to the way that biases in a learning procedure affect frequency in iterated learning. I make two crucial advances beyond this conception of evaluation and change, however. First, change is explicitly modeled. We cannot assume *a priori* that the circumstances of learning are such that any difference in a metric generates a difference in the predicted probabilistic typology. Thus it is important to consider learning and its iteration as part of the model of typological prediction, not incidental to it. Second, the metric is not stipulated for its own purposes in my approach. There is no explicit goal of simplicity—the act of attempting to learn *necessarily* creates a bias towards patterns which are easier to learn.

The approach to typology developed in this dissertation is in a sense quite traditional, grounded in the same ideas as the evaluation metric. However, my metric is induced automatically through the combination of a learning theory and representational system and learning is explicitly modeled.

1.4 Basic stress tendencies

We may now turn to the typological data of interest, before moving on to existing models and the learning approach proposed here. The two sources for my typological counts are StressTyp (Goedemans, 2010) and the Stress Pattern Database (Heinz, 2007). These databases serve as a useful summary of the typology and are readily searchable. However, they are not without potential weaknesses for this kind of numerical work. Neither database is balanced, either for area of the world or genetic affiliation. Thus some areas are much better represented in the databases, while others are underrepresented. The same holds true for language families and subgroupings that have been studied more or less extensively. This lack of balance contrasts with databases such as the UCLA Phonological Segment Inventory Database (Maddieson, 1980), for which considerable effort was expended to ensure some degree of equal representation. To address this point, in this dissertation I focus principally on numerical biases which appear robust. These asymmetries are the ones which are least likely to be due to missampling the extant (or once extant) languages of the world. In Appendix A, I show that the generalizations of interest in this work largely stand up to several types of random resampling based on genetic affiliation.

I begin with a discussion of the typology of fixed stress. Fixed stress systems are those in which stress always falls a given distance from a word edge (if possible, given word length). Unless otherwise mentioned, by this I intend systems with a *single* fixed stress. Thus there are final stress systems in which stress always falls on the final syllable (distance = 0), penultimate stress systems where it falls on the second-to-last (in words of two or more syllables, distance = 1), and antepenultimate stress systems where stress falls a syllable farther away still (distance = 2). On analogy, initial, peninitial, and postpeninitial systems are described.

All six of these patterns have been reported in some capacity, although the place of postpeninitial stress at least is greatly contested (Hyman, 1977; Gordon, 2002). When

	From left	From right
	σσσσσσσσ́	*σσσσσσσ́*
Distance 0	initial	final
	69 languages	74 languages
	σσ́σσσσσ	*σσσσσσ́σ*
Distance 1	peninitial	penultimate
	12 languages	60 languages
	σσσ́σσσσ	*σσσσσ́σσ*
Distance 2	postpeninitial	antepenultimate
	0 languages	8 languages

Table 1.3: Fixed stress languages with counts from Heinz's (2007) Stress Pattern Database

	From left	From right
'Distance 0	52 languages	91 languages
Distance 1	26 languages	46 languages
Distance 2	3 languages	5 languages

Table 1.4: Rounded expected counts from a χ^2 test of Table 1.3: $\chi^2 = 25.39$, $df = 2$, $p < 0.05$.

the word-edge distance is small, these are some of the most common stress patterns across languages—initial, final, and penultimate stress are very well attested.

Several asymmetries are readily apparent in the probabilistic typology (Table 1.3). These are in part supported by a significant χ^2 test ($\chi^2 = 25.39$, $p < 0.05$), suggesting the non-independence of distance and edge. First, patterns with a distance of higher than 1 from the word edge are rare. That is, antepenultimate and postpeninitial stress appear markedly distinct from the four other patterns. This type of asymmetry is mirrored in the assertions that the categorical typology does not contain postpeninitial stress. Antepenultimate stress, however, is common enough to clearly warrant inclusion.

In addition to the rarity of distance 2 patterns, distance 1 patterns show interesting tendencies. Peninitial stress is distinctly rarer than initial stress, but this difference is far less on the right edge of the word. This can be seen in the expected counts from

the χ^2 test—penultimate stress is more common than expected, while final stress is less common (Table 1.4, discussed further in Chapter 3). Despite this, there is an apparent distinction between distance 0 patterns and distance 1. Combined with the previous generalization, we can thus see that for fixed stress systems frequency decreases as distance from the word edge increases.

Another asymmetry relates to the edge referred to by the stress system. Left edge systems appear rarer overall—apart from initial stress, we have only the rare peninitial systems and marginal postpeninitial ones. This asymmetry could break down in three logically distinct ways. First, peninitial stress and postpeninitial stress could be comparatively disadvantaged, with penultimate and final stress at the "baseline" for systems without a substantive bias. Second, penultimate and final stress could be comparatively *advantaged*, with peninitial and postpeninitial at baseline. Finally, it could be that the two pairs are advantaged *and* disadvantaged. I return to this left/right asymmetry in Chapter 3.

1.4.1 Iterative stress typology

Of course, stress need not be fixed with respect to word edges. There are also so-called iterative systems. These are stress patterns in which stress occurs at intervals from the edge or from the main stress. Here I only explicitly discuss patterns which make no reference to syllable weight (quantity insensitive) or otherwise to lexically-indicated stress. These systems will productively treat strings of equal length, measured in syllables, as identical for stress purposes.

Lexical exceptions to productive stress systems are very common. Here I mean to exclude systems where the *typical* mode of stress assignment derives from a lexical indication of the placement of stress. It is difficult to precisely specify "typical" stress patterns—here I rely on the primary description given in stress typologies.

	No degenerate feet (Binary)	Degenerate feet (Nonbinary)
Trochaic	$(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma})$
Iambic	$(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\sigma\acute{\sigma})\sigma$	$(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\acute{\sigma})$

Table 1.5: Parametric left-to-right patterns

Foot type	Direction	Degenerate feet?	Count
Trochees	Left-to-right	no	33
		yes	22
	Right-to-left	no	34
		yes	4
Iambs	Left-to-right	no	13
		yes	3
	Right-to-left	no	2
		yes	3

Table 1.6: Parametric iterative stress in StressTyp. *Degenerate feet?* indicates whether all feet are binary. That is, "no" indicates that degenerate feet are not permitted.

The presence or absence of lexical stress is ultimately not crucial if the tendencies modeled do not reflect lexical stress.

These iterative patterns can be looked at in a number of ways. One illuminating approach is to view them is in terms of their necessary structure when parsed using metrical feet (e.g. Hayes, 1995). That is, strings can be parsed into trochees (left-headed feet), iambs (right-headed feet), or degenerates (monosyllabic feet). This terminology gives a set of parameters with which to divide up small predicted typologies.

Perhaps the most striking asymmetry in the probabilistic typology is the preponderance of trochees over iambs. This crosslinguistic preference for trochaic patterns over iambic ones has been noted many times (e.g. by Hayes, 1995). I return to this preference in Chapter 3.

$\acute{\sigma}$
$\acute{\sigma}\sigma$
$\acute{\sigma}\acute{\sigma}\sigma$
$\acute{\sigma}\sigma\acute{\sigma}\sigma$
$\acute{\sigma}\acute{\sigma}\sigma\acute{\sigma}\sigma$
$\acute{\sigma}\sigma\acute{\sigma}\sigma\acute{\sigma}\sigma$

Table 1.7: Example strings in the initial clash language

Among the trochaic patterns there is a clear split between the first three in Figure 1.6 and the last pattern. That is, right-to-left trochaic parses with degenerate feet are far less common than would be expected otherwise given the frequency of other trochaic parses. This type of language is unique among these eight in that it contains an initial sequence of stressed syllables in odd-parity words (an initial clash). Languages with an edge clash will be of continued relevance in Chapter 2.

Due to their overall smaller numbers in the sample, it is difficult to accurately classify the asymmetries within the iambic languages. Previous work has taken the third iambic language (*the initial lapse language*) to be unattested (Alber, 2005), while the others are argued to exist in some form. It is also reasonably clear that left-to-right iambs without degenerate feet are probably the most common among these iambic patterns. As my summary of the probabilistic typology, I take this language to be "common" for iambs, the initial lapse language to be "very uncommon" (i.e. unattested?), and the remaining two to be "uncommon."

1.4.2 Previous accounts of the asymmetries

1.4.2.1 Fixed stress

Equipped with typological data, we may now consider existing attempts at explanations of frequency, beginning again with fixed stress. Peninitial stress is considerably less common than penultimate stress. This asymmetrical attestation has previously been attributed to the effect of nonfinality on typology (Hyman, 1977;

Gordon, 2000). In these accounts, nonfinality effects in typology arise due to the predominance of final boundary tones across languages. With final boundary tones, the account goes, final stress becomes more difficult to distinguish. Thus stress is pushed off final position onto the preceding syllable. No such pressure exists yielding peninitial stress, however. Furthermore, initial boundary tones are dramatically less common across languages (Gordon, 2000). Therefore, any "noninitiality" effect that *is* imposed upon typology is much weaker. This approach generalizes to more word-internal stress patterns. Antepenultimate stress could conceivably result from right edge avoidance, but no such account would be generally available for postpeninitial stress.

These accounts are not necessarily tied to the typology of boundary tones, however. Any phonetic pressure to avoid final stress could have a similar effect on typology. Thus accounts of nonfinality linked to, for example, final lengthening (e.g. Lunden, 2006) could be equally successful. I return to these accounts of nonfinality as perceptual effects on typology in Chapter 3.

It is not true that there are *no* conceivable motivations for peninitial stress. Such a pattern could follow from, for example, a pressure for consistency in stress patterns. Peninitial stress could be consistently represented as a left-aligned, disyllabic iamb. It is striking, then, that such representational or analytic effects so infrequently overcome the apparent phonetic asymmetry.

This type of approach to typological asymmetry in fixed stress, based on an asymmetrical pressure against final stress that does not apply to initial stress, provides an explanation for the disparity between peninitial and penultimate stress. It seemingly does not account for the general disparity between right-counting systems and left-counting systems. In Heinz's (2007) Stress Pattern Database, final, penultimate, and antepenultimate stress account for 142 languages while initial and peninitial only account for 81. A preference for right-counting stress does not follow from nonfinality

in the absence of further elaboration. Put another way, penultimate stress is overattested not just in comparison to peninitial stress but in relation to the (non-)difference between initial and final stress.

1.4.2.2 Position of clash/lapse

As noted above, systems with clash or lapse at the left edge are less common than systems with clash or lapse at the right edge. Kager (2001) addresses the asymmetry in lapses by positing the constraint Lapse-At-End. This constraint is violated by lapses which occur in any position other than the right edge. This results in systems which will not tolerate lapse in general, but will tolerate it in this one circumstance. An example of such a language is the final lapse language—left-to-right trochees with no degenerate feet. His account additionally renders the initial-lapse languages impossible—there is no constraint motivating *only* an initial lapse, so a language that tolerates lapse in just this position is not generated.

This account does not make any claims about the position of clash, however. This contrasts with the typology reviewed previously—initial clash is strongly disfavored relative to other trochaic patterns. This account cannot be a full explanation of the relative frequency of positions for clash and lapse.

1.4.2.3 Foot shape asymmetry

It is sometimes claimed that trochees are universally preferred over iambs, supported by acquisition and learning data (e.g. Jusczyk et al., 1993; Adam and Bat-El, 2009). This would potentially account for the bias for trochaic systems in the probabilistic typology—trochees are preferred because they are learned more easily or because the acquisitional system tends to posit them instead of iambs.

This account is called into question by conflicting results in acquisition (Vihman et al., 1998). One possibility is that the apparent experimental preference for trochees over iambs may instead result from trochaic regularities in the learning data. That

is, trochaic bias might be evident in such experiments because typical sorts of data learners encounter are more compatible with the phonetics of trochees, rather than reflecting a true bias in favor of a trochaic. Another possibility is that if there are already more trochaic languages than iambic ones, and if experiments simply show a bias for the dominant foot type of the ambient language, more of these experiments will show trochaic bias simply because there are more trochaic languages. That is, we may see more experiments reflecting trochaic bias than iambic bias because these experiments are essentially sampling the properties of subjects' early learning. If more languages in the world are trochaic (or, at least, languages with more speakers), this creates a tendency for more subjects to be exposed to trochaic-type data early in life. A sampling process that exposes this merely echoes the statistics of the typology, it does not fundamentally illuminate its origins. It is thus at least questionable whether a universal preference for trochees as a foot shape is desirable. Another type of pressure may instead be required.

1.4.3 Typological correlates

Bane and Riggle (2008) identify some *typological correlates* of the probabilistic typology of quantity-insensitive stress. In particular, the three they discuss are: the trigram entropy of stress patterns, stress systems' confusability with other systems, and the number of constraint rankings which produce these patterns.

The first correlate, trigram entropy, can be thought of as measuring the regularity of a stress pattern. Entropy is a measure of the information conveyed by a random variable. Random variables represent properties of systems taken to follow probability distributions—in this case, the choice of trigram in a stress system. It is maximized when the variable is uniformly distributed over its possible values. It is for a uniform distribution that we may predict the smallest amount of information about a variable ahead of time—we have the least amount of knowledge about the value a

random variable will assume. Thus, the variable *itself* conveys the highest amount of information—has the highest entropy. The trigram entropy specifically measures the entropy of trigrams—sequences of three syllables (or word edges) for the case of quantitative stress. When trigram entropy is high, less can be known ahead of time about a particular sequence of three syllables (or word edges) than is possible. Thus, patterns with high entropy are patterns that have many different trigrams in them evenly distributed, rather than just a few occurring most of the time.

Bane and Riggle find that trigram entropy is negatively correlated with the frequency of a stress pattern. That is, the less variable the trigrams of a pattern are, the more likely that pattern is to be well-attested. This makes intuitive sense from the perspective of learning a pattern. If a pattern has reliable trigrams, the learner can use these few prototypical trigrams as a model of the actual pattern with substantial success. If the trigrams are less reliable—more entropic—however, this strategy will take more evidence to overcome this entropy.

Their second correlate is confusability. By this they mean the minimum length of the strings needed to disambiguate a pattern from all other patterns. For example, it is insufficient to observe only strings of under four syllables to learn antepenultimate stress—without four-syllable forms such a pattern looks identical to (for example) initial stress. Bane and Riggle find that the less "confusable" a pattern is, in this sense (that is, how short the forms needed are), the more well-attested a pattern is likely to be. Again, this is understandable from a perspective of learning. Longer forms tend to be less common than shorter ones, so if longer forms are *required* to isolate a particular pattern, we should expect such a pattern to be less quickly learned. This correlate is interesting compared to the first one because it is theory-dependent. The patterns with which a set of strings are consistent are determined by the set of *possible* patterns. Thus across theories of categorical typology we might find differences in confusability measures for particular patterns. In general, however,

the correlation between this measure applied across theories would likely be strong. Bane and Riggle find, in fact, that it is a fruitful metric for more than one model of the categorical typology.

Bane and Riggle's final correlate of relative attestation is the number of rankings which describe a pattern in an Optimality Theoretic model. That is, the more ways there are of describing a pattern via ranking, the better attested that language is. This approach follows work by Coetzee (2002) examining the link between typological frequency and the number of rankings describing a pattern. This correlate is less inherently linked to learning. It would follow if, for example, learning consisted of randomly sampling from the space of consistent rankings under Optimality Theory. Riggle (2008) proposes a learning algorithm similar to this idea, biasing the learner toward results with many possible rankings. However, this measure is also likely correlated with others. The work of creating OT constraint sets proceeds by identifying regularities in typology. Further, it is preferred to reuse constraints in analyses rather than posit new ones. This means that OT constraint sets proposed by analysts have general pressures on them pushing them towards generality and large applicability of constraints. These sorts of considerations *do* seem possibly useful for language learning—the number of valid grammars producing a language has potential impact even in non-OT theories.

Bane and Riggle's contributions to an understanding of the probabilistic typology of quantity-insensitive stress are quite valuable. Their correlates—particularly the first two—are important to the performance of the learner I present. I formalize these ideas within an explicit learning model and extend them in several ways.

1.5 Statistical regularity, learning, and typology

1.5.1 Constraint assumptions

Now I return to the types of languages discussed above: fixed stress and basic iterative stress. I model asymmetrical attestation among these patterns with learnability differences emerging from an online learner of MaxEnt grammar. In this section I assume a constraint set designed to be relatively uninformed by typology. Thus these constraints do not reflect the insight of analysts, but instead are a general-purpose way to attempt to model strings of the type found in stress.

Each constraint refers to an n-gram, or a sequence of n adjacent elements. The elements are of three types: stressed syllables, unstressed syllables, and word edges. n may be 1, 2, or 3. Every constraint matching this template is included in the model—no asymmetries are introduced prior to learning.

An example unigram constraint is $*\sigma$, a constraint that penalizes unstressed syllables. This is somewhat analogous to the more standard Parse(Syll). One bigram is $*\acute{\sigma}\#$, penalizing stress at the end of a word like Nonfin(Syll). Trigram constraints are typically more exotic, for example $*\#\acute{\sigma}\acute{\sigma}$, penalizing initial clash.

1.5.2 Emergence of the perfect grid

Another useful way of categorizing iterative stress systems is in terms of whether or not they are *perfect grids* (Prince, 1983). Perfect grids are patterns in which there are neither clashes nor lapses—every word length is characterized by a stressed or stressless initial syllable followed by perfect alternation between the two. Given the assumed constraint set, such systems are more distinct and more learnable.

In the empirical typology, we see that the perfect grid languages are relatively common within a foot type. All perfect grids are common, with the exception of the right-to-left iambic languages that tolerate degenerate feet.

Foot type	Direction	Degenerate feet?	Perfect grid?	Count
Trochees	Left-to-Right	no	no (right lapse)	33
		yes	yes	22
	Right-to-Left	no	yes	34
		yes	no (left clash)	4
Iambs	Left-to-Right	no	yes	13
		yes	no (right clash)	3
	Right-to-Left	no	no (left lapse)	2
		yes	yes	3

Table 1.8: Iterative stress typology as perfect grids

This asymmetry follows from the statistics of the learning problem given the model presented here. Perfect grid languages are just those in which the necessary constraints are most reliable. That is, the weightings that are needed in order to produce perfect grid patterns are weightings in which the highly-weighted constraints have consistent, low violations. This means that the strings within a language are distinct from those outside of it: candidates in the language share this consistency, while candidates outside of it do not have such consistency.

To see this, we can start by working backwards. If a language's necessary constraints are very reliable in this sense, making it distinct, it must not require constraints which make reference to more than two positions. This is the case because no trigram (or larger n-gram) is reliably followed in short words. Any trigram pattern true of long words will be necessarily interrupted when words are short enough that both word edges are within syllable-adjacency of falling into a trigram. That is, disyllables typically break prevailing patterns, measured in a trigram sense. Unigram constraints make overly strong demands when used alone. There are only three possible types of stress languages which fully obey unigram constraints: the fully unstressed language type (*$\acute{\sigma}$), the fully stressed language type (*σ), and the empty language type (*#). We must therefore focus on languages which require only bigram

Pattern	Frequency (Heinz, 2007)
initial stress	69
final stress	74
initial and final stress, clash in disyllables	3
initial and final stress, iambic disyllables	1
initial and final stress, trochaic disyllables	0

Table 1.9: Bigram patterns and their frequency

constraints. This focus on bigrams emerges from statistical relationships, not from an assumption focusing on bigram constraints.

Languages that can be represented by bigram constraints alone must either be non-iterative or be perfect grids. The non-iterative languages are those with only edge stress. There are five such language types, as shown in Table 1.9. The typology of languages representable with these n-gram constraints was validated in OT-Help 2 (Staubs, et al. 2010). This is software which allows typology calculations in parallel and serial OT and HG based on specified patterns of violation.

The latter three language types suffer from less reliable constraints due (once again) to disyllables. The configurations called for by such languages in longer words do not match with those in such short words (as highlighted in the descriptions). For example, the third language type will not tolerate clash except in the case of disyllables. This unreliability is reflected in a slower learning rate for these languages, as shown in Figure 1.3. This figure gives the residual error for a learner learning a particular language type after a given number of iterations. Error is measured as sum squared error—the sum of the squared difference between the probability the learner gives to a candidate and the probability given by the teacher.

As for the iterative languages, they may not contain any clash or lapse. This is because such strings are "bigram-ambiguous" with other strings (Table 1.10). That is, their violation profiles are exactly the same on all bigram constraints.

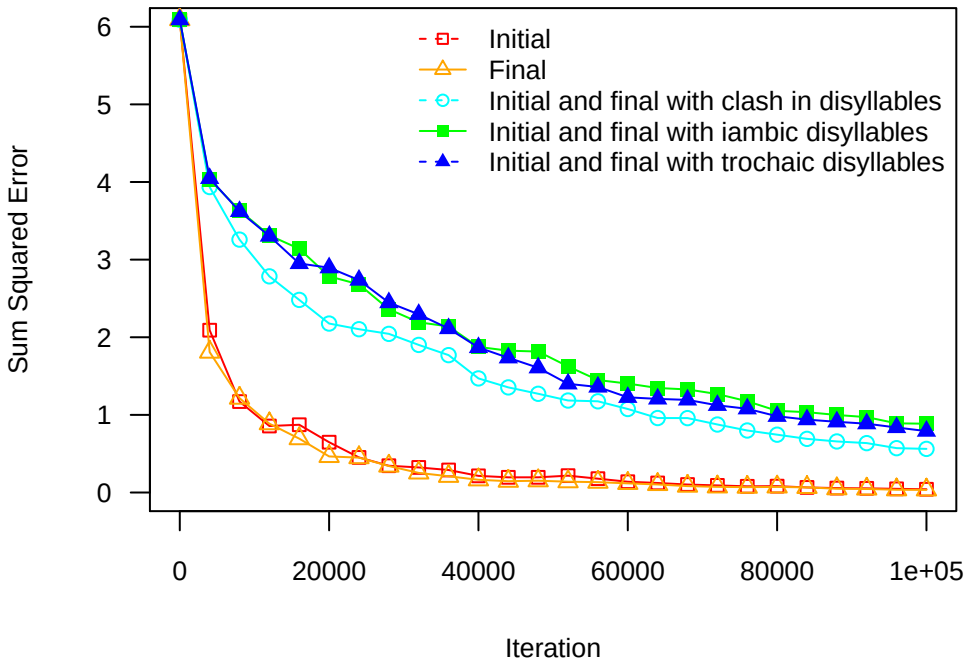


Figure 1.3: Bigram patterns. Initial and final stress are more readily learned due to their consistent patterns. Results are the average of 10 trials. $\eta = 0.1$.

$\#\acute{o}\sigma\acute{o}\sigma\acute{o}\sigma\sigma\#$
$\#\acute{o}\sigma\acute{o}\sigma\sigma\acute{o}\sigma\#$
$\#\acute{o}\sigma\sigma\acute{o}\sigma\acute{o}\sigma\#$

Table 1.10: Example of bigram-equivalent candidates for 7-syllable words with one lapse. Bigrams allow enforcing the existence of one and only one lapse, but not its position.

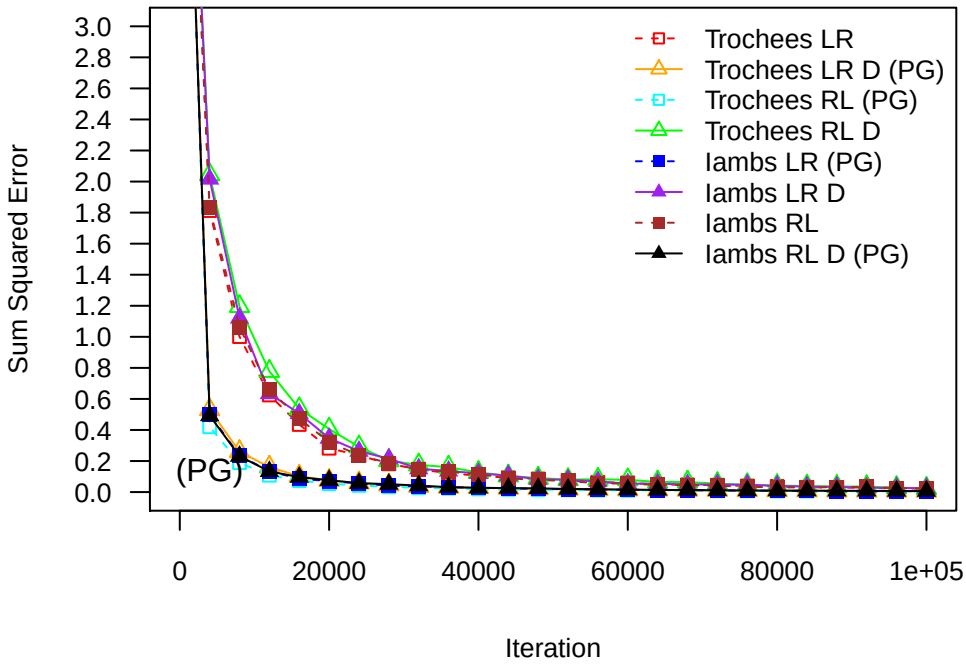


Figure 1.4: Learning perfect grids. With unrestricted bigram and trigram constraints, perfect grids are learned faster. Perfect grid patterns indicated with (PG). Results are the average of 10 trials. $\eta = 0.1$.

The learner is tasked with choosing one and only one of these bigram-ambiguous strings as optimal, which is clearly impossible with only bigram constraints. For these languages the learner must make use of constraints which are larger than bigrams (*viz.* trigrams). These constraints *are* available to the learner, but as already discussed, only bigram constraints maximize constraint reliability and language distinctiveness. Due to their distinct candidates, perfect grid languages are learned more readily than any other iterative patterns. This is shown in Figure 1.4.

1.5.2.1 Full bigram typology

The summary of bigram typology above is limited in two ways. First, the Harmonic Grammar-type evaluation in MaxEnt introduces additional possibilities beyond the Optimality Theoretic factorial typology. Second, the typology includes languages which "over-stress." I will discuss each of these in turn.

There are in total 22 languages describable with only bigram constraints under Optimality Theory. First there are the 9 languages described above:

1. 4 perfect grid languages

2. 5 fixed stress languages

The more complete OT typology includes a language without stress and a fully-stressed language (the unigram-compliant languages mentioned). Both of these languages can be seen as essentially equivalent: they are languages in which stress uniform over a word. Typically culminativity is taken to be universal—at least one syllable is "more stressed" in every word of a language (e.g. Hayes, 1995). If this universal is needed with such a constraint set, these languages seem to require (and compel) exclusion on functional grounds or through implementing a filter on the output of GEN or EVAL. This is a useful reminder of the potential pathologies lurking in even very simple OT constraint sets, as well as a demonstration that typological prediction based on only one source of typological structure is likely to make faulty predictions.

The 5 fixed stress languages each have "inverse" patterns. These patterns are exact duplications of the 5, with stressed syllables switched for unstressed ones. Thus there is a pattern in which all but the last syllable is stressed, all but the edges, and so on. These languages are problematic in that they seem to "over-stress." That is, more syllables are stressed than we will typically see across languages. This is perhaps for functional reasons—for example, stress may be less perceptible in succession. Perhaps they could additionally be proscribed along the lines of a filter as mentioned above, or perhaps their exclusion results entirely from phonetic pressures. This sort of pathology is avoided in part if foot structure is used—non-minimal feet create domains in which stress alternation is mandatory. I will not discuss this issue further, though I return to the issue of similar extragrammatical pressures in Chapter 3. In addition,

initial, final, and stressless languages (and their inverses) each have variants differing only in whether monosyllables are stressed. These variants are likely typologically unimportant.

Harmonic Grammar evaluation adds 30 additional languages to the predicted typology. The majority of these (19 of the 30) contain stressless strings. If these languages are categorically proscribed on other grounds for OT evaluation, these are not an additional issue.

The remaining languages are all fully-stressed languages up to a certain word length, then some fixed-stress pattern thereafter. These languages then fall under the same sort of considerations owed to the "inverse" languages.

Some of these categorical predictions of the n-gram constraints point to problems with this choice of constraint set. I emphasize that this choice is made here for simplicity to demonstrate the kinds of biases that can emerge from very simple assumptions. Many of these results carry over to more refined theories of Con, through similarities between n-grams and the ways e.g. clash and lapse are reckoned. I show results with these sorts of constraint sets in Chapters 2 and 3.

1.5.3 Fixed stress

As discussed in (§1.4), fixed stress systems show a bias towards stress closer to the word edge. For example, penultimate stress is much more common than antepenultimate stress. This trend follows from the same considerations as the emergence of the perfect grid.

Apart from perfect grids, final and initial stress systems come closest to maximizing constraint reliability and language distinctiveness. This is not true, however, of fixed stress systems which place stress farther inside a word. These languages suffer from unreliability caused by small words. For example, penultimate stress can be confused with initial stress analyses in disyllables—or with peninitial stress in trisyl-

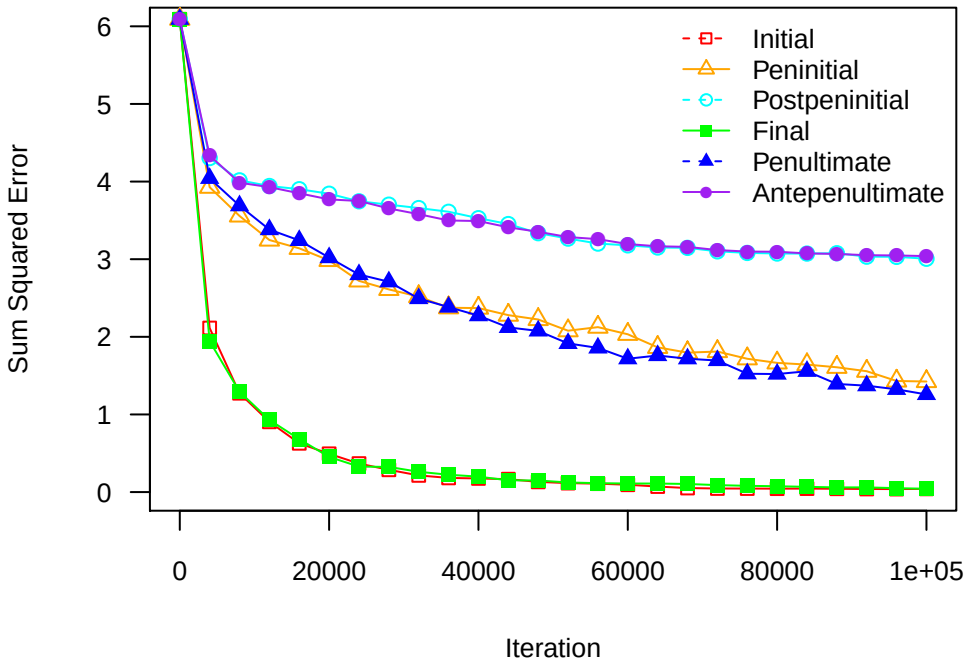


Figure 1.5: Fixed stress. The closer fixed stress is to a word edge, the faster it is learned. Results are the average of 10 trials. $\eta = 0.1$.

lables. This confusability increases as distance from the word edge increases, yielding a decrease in learning rate. Thus the empirical typology for this pattern is mirrored by the predicted typology: fixed stress, generally speaking, becomes less common as distance from the word edge increases. The learnability distinction made between each language is shown in Figure 1.5.

1.6 Learning bias

1.6.1 Bias from distinctiveness

As discussed above, any learning algorithm (along with its representational space) is biased for and against certain classes of patterns. In the previous section, I used this fact to demonstrate that learnability differences can be usefully associated with differences of relative attestation. However, one might reasonably be left wondering why these learnability differences exist. For models with MaxEnt grammars learned

with SGA, it is necessary to understand how these kinds of systems can be made to learn faster or slower. In iterated learning, these differences partially account for accumulated bias in the predicted typology. The particular biases of this system are fairly general, applying also to a number of other approaches to error-driven learning.

To understand the biases of SGA, we must consider violation vectors more carefully. Recall that violation vectors are the ordered sets of violations incurred by candidates on each constraint. For example, consider a candidate [badat] with a constraint set consisting of *VoiceObs, *Coda, and *[velar], militating against voiced obstruents, codas, and velars, respectively. This candidate would have 2 violations of the first constraint, 1 of the second, and none of the third. Its violation vector is therefore $\langle 2, 1, 0 \rangle$.

Violation vectors are the true objects of representation in Optimality Theory-like learning. They share this property with many or most approaches in machine learning, where feature vectors are used to represent instances of interest. In OT, violation vectors are treated as primary throughout learning work (e.g. Tesar and Smolensky, 2000), but are only seldomly discussed as the primary representation of candidates themselves (but see Golston, 1996). Despite this, violation vectors are implicitly the representational mechanism throughout Optimality Theoretic work. Questions of candidate representation are always fundamentally questions of their representation as violation vectors, whether this be the inclusion or exclusion of a constraint in Con, a question of how a constraint assesses violations, or so on.

Violation vectors are important for understanding learning biases because they establish the geometry of the learning problem. Violation vectors exist in a n-dimensional space, where n is the number of constraints. In this space, some vectors are more similar to others while some are more distinct. These types of relationships are potentially of great importance to learning.

In particular, in the error-based SGA it matters how distinct vectors within a language are from those outside it. This is an intuitively attractive notion: the process of learning a language in an Optimality Theoretic type of grammar is essentially the process of separating out "good" candidates (winners) from "bad" ones (losers). If the winners look quite different from losers, the task of separating them grows easier.

A classic result by Novikoff (1962) gives a convergence guarantee for perceptron learning when applied to linearly-separable classes. This result shows a link between the speed of perceptron learning and properties of the vectors considered for classification. Learning speed decreases when the norm (roughly, the size) of the largest vector under consideration increases. Speed *increases* when the margin between vectors in one class and those in another class increases.

Such classes are like those found in HG—in every tableau (at least), a hyperplane can be drawn in the space of violation vectors to separate out the winners from the losers. The perceptron algorithm is essentially identical to SGA for MaxEnt grammar. Thus these two considerations apply simply to error-driven learning of constraint-based grammars. The size of the largest vector considered corresponds roughly to the most-violating candidate that must be considered in learning. That is, the wider the range of candidates that need to be ruled out or accepted, the slower that learning proceeds. The margin between vectors corresponds to the distinctiveness of the candidates which are part of a language from those which are excluded from it. Thus the speed of learning increases when candidates in a language are all similar to one another (in terms of violations vectors) in ways in which they are dissimilar to candidates out of the language. A margin of separation of this type has a role in the optimization work of Potts et al. (2010) and the adapted perceptron convergence proof for Noisy HG of Boersma and Pater (2014).

The connection between violation vectors and learning is crucial. Learning results will necessarily depend on assumptions about the constraint set, precisely because

the only way a single linguistic pattern can differ from another is through their representations in violation vector space. In fact, though a learning bias may hold over a large number of assumptions, it is unlikely to hold over all possible representations. Examples of this failure of generalization are given in later sections.

1.6.2 Distinctiveness proof sketch

In this section I extend the intuition built above into a more formal intuition for the mechanics behind a distinctiveness bias. We are interested particularly in the question of how the structure of the violation vectors $\mathbf{v}_i$ affects the practical learning rates attainable by MaxEnt learners using algorithms like SGA. The first question for such an investigation should be whether SGA is necessarily learning at all, and what this means. For fully-observed data (no hidden structure), SGA is guaranteed to converge on a weight vector $\mathbf{w}$ (Jäger, 2007) which chooses the target language as optimal in the HG sense, assuming the language is in fact describable by an HG grammar. Beyond this, the algorithm will also successfully increase weights such that the probabilities of target forms within tableaux arbitrarily approach 1. SGA with Robust Interpretive Parsing will encounter local optima (as in the Noisy Harmonic Grammar simulations of Boersma and Pater, 2014)), but will improve its error with respect to the starting position, assuming that position is not itself a local optimum.

We know that these learning algorithms do learn, and thus the relevant question is instead how $\mathbf{v}$ affects how *quickly* the algorithm learns. We may look at this in several ways. Perhaps the simplest approach is to examine the residual error of the learner after some "sufficient" amount of learning. For example, the sum squared error (SSE), SSE^t, measures the degree of divergence between the learner's hypothesis and the target language at iteration t. It is calculated as the sum of all the squared differences between the learner's probabilities for a set of candidates and the observed probabilities.

$$\mathrm{SSE}^t = (\mathbf{p}^t - \mathbf{p}^*)^T(\mathbf{p}^t - \mathbf{p}^*) \tag{1.10}$$

$$= \sum_{i \in \textsc{Inputs}} \sum_{j \in \textsc{Gen}(i)} \left(p^t(j|i) - p^*(j|i)\right)^2 \tag{1.11}$$

The error SSE^0 at the initial weights $\mathbf{w}^0$ is not due to learning. It comes about purely due to the distribution of the starting state. Thus to model the effect of learning alone, it is best to look instead at the reduction of error between that starting state and time t.

$$\Delta\mathrm{SSE}^{0t} = \Delta\mathrm{SSE}^0 - \Delta\mathrm{SSE}^t \tag{1.12}$$

This change in error between the start and iteration t is entirely described by the changes in error between each iteration of learning.

$$\Delta\mathrm{SSE}^{0t} = \sum_{k=1}^{t} \mathrm{SSE}^{(k-1)(k)} \tag{1.13}$$

A full analysis of learning requires more than examination of $\mathrm{SSE}^{(k-1)(k)}$, because these changes in errors are strongly correlated across iterations. That is, the hypothesis at one iteration affects the possibilities for error reduction in the next update. Despite this, an inspection of a single change can build a strong intuition for the mechanics of learning speed in these models.

Each $\Delta\mathrm{SSE}^{0t}$ can be characterized in terms of a set of changes to candidate probabilities $\Delta\mathbf{p}$ and thus a change in weights $\Delta\mathbf{w}$.

$$\Delta\text{SSE}^{(k-1)(k)} = \sum\sum \left(p^k(j|i) - p^*(j|i)\right)^2 - \sum\sum \left(p^{k-1}(j|i) - p^*(j|i)\right)^2 \quad (1.14)$$

$$= \sum\sum \left(p^k(j|i) - p^*(j|i)\right)^2 - \left(p^{k-1}(j|i) - p^*(j|i)\right)^2 \quad (1.15)$$

For fastest learning, this quantity should be minimized. The minimum value is obtained with a p^k that is as close as possible to p^* and a $p^{(k-1)}$ that is as far away as possible. That is, a single learning step is most effective in an error-reduction sense if the probabilities change a large amount. The change in probabilities $\Delta p_{ij}^{(k-1)(k)}$ can be defined from the MaxEnt definition of candidate probability.

$$\left|\Delta p_{ij}^{(k-1)(k)}\right| = \left|p_{ij}^k - p_{ij}^{k-1}\right| \quad (1.16)$$

$$= \left| \frac{e^{(\mathbf{w}^k)^T \mathbf{v}_{ij}}}{\sum_x e^{(\mathbf{w}^k)^T \mathbf{v}_{ix}}} - \frac{e^{(\mathbf{w}^{k-1})^T \mathbf{v}_{ij}}}{\sum_x e^{(\mathbf{w}^{k-1})^T \mathbf{v}_{ix}}} \right| \quad (1.17)$$

Changes in weight are associated with absolute changes in probability. Some weight changes do not make substantial difference to probabilities: for example, a weight that has already pushed a probability to a logistic asymptote will not affect probabilities much if it is changed only slightly. In general, however, greater changes in weights mean greater changes in probability. This is especially the case "early" in learning, when the learner's probabilities differ radically from its teacher's.

Therefore, one way to maximize the effectiveness of early learning (sum squared error) is to maximize the (norm) change in weights. Recall the update rule in SGA:

$$\text{New Weights} = \text{Old Weights} + \eta \times (\text{Teacher Violations} - \text{Learner Viol.}) \quad (1.18)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta(\mathbf{v}_i^* - \mathbf{v}_{ij}) \quad (1.19)$$

To maximize the change between $\mathbf{w}^t$ and $\mathbf{w}^{t+1}$, the quantity $\mathbf{v}_i^* - \mathbf{v}_{ij}$ should be maximized. This ensures that the difference in the update rule is large and thus that the weights are changed to a greater degree.

Restating this, the teacher's form should be as far apart from the learner's choice as possible. That is, the learner and teacher should have *maximally distinct* optima in early learning. Any candidate chosen by the learner and not by the teacher is necessarily outside the target language. Therefore, approximating from the single step case, initial learning in SGA is fastest when the distance between violation vectors within a target language are far apart from those outside it.

1.6.3 Redundancy of representations

The type of learnability bias discussed informally for stress is not the only one that emerges from the models I propose. Another sort of bias is based on the relative redundancy of representations, something of intuitive use for analysts. Despite a relative lack of concern for frequency, researchers do occasionally speak to the implications of theories for frequency. The issue with this type of discussion is that it fundamentally relies on intuition about the relationship between a theory and its frequency implications: Optimality Theory (for example) comes with no *inherent* way in which to derive typological frequencies. Therefore, any argument from frequency based solely in OT will be, in some sense, without basis. This is not to say that the arguments themselves are groundless once a frequency theory is adopted, however; approaches based in learnability or r-volume can have great accord with analysts' intuitions.

As a motivating example outside the domain of stress, I consider Padgett's (1995) discussion of color harmony. In harmony systems, it does not appear that back harmony and round harmony are independent. Instead, it seems that there is a

strong positive correlation between the presence of one process and the presence of the other, especially in Turkic languages.

The class "color" is intended to capture the generalization that backness and rounding are not fully independent. If there were no special connection between the features, we should expect harmony for either, both, or neither, with no necessary expectation on frequency. This is contrary to the apparent fact that the systematic agreement of two vowels on backness increases the likelihood of their agreement on rounding, and *vice versa*. The intuition behind the class "color" is that perhaps these two features often track together because they are both aspects of one larger class encompassing them both (while still excluding others). Padgett rightly points out that most existing theories do not account for the relationship between the individual features backness and roundness.

> In spite of a fair precedent for a class *Color*, virtually all researchers in the generative tradition addressing Turkish vowel harmony have assumed two separate rules/constraints of harmony, implicitly or explicitly: one harmony of [back], and one of [round], a state of affairs rendering color harmony as likely seeming as a co-patterning of [back] and [nasal]. (Padgett, 1995, p. 390)

The issue with this statement is that it goes further than pointing out prior accounts do not deal with this relationship, claiming that most of these theories predict co-patterning of back and round to be similar to co-patterning of back and nasal. In fact, most theories say *nothing at all* about the comparative frequencies of these two sets of languages, provided that both are predicted to occur at all. This is because standard generative phonology has no formal means to make distinctions in frequency.

Nascent within this view, however, is a proposal for a true account of a frequency tendency. The abstraction "color" groups backness and roundness in a way not available to, for example, nasality paired with ATR or backness. This asymmetry in representation can create an asymmetry in learnability, biasing results in favor of harmony systems which obey color generalizations.

I demonstrate this type of learning result with a toy representation of harmony systems. 100 words are randomly generated with two vowels, chosen uniformly at random. Of these vowels, some number are chosen to be underlyingly disharmonic according to four types of harmony: A, B, C, and D. A "disharmony" with two vowels is simply any mismatch in feature values between the two. These candidates are represented by six constraints, principally encoding the available types of disharmony:

1. Harmonize(A): Assign a violation for every disharmony of type A in a word.
2. Harmonize(B): Assign a violation for every disharmony of type B in a word.
3. Harmonize(C): Assign a violation for every disharmony of type C in a word.
4. Harmonize(D): Assign a violation for every disharmony of type D in a word.
5. Harmonize(AB): Assign a violation for every disharmony of type A or B in a word.
6. Ident: Assign a violation for every vowel feature that is mismatched between input and output.

Note crucially the presence of Harmonize(AB) and the absence of Harmonize(CD). This echoes the opposition of a class for color and an absence of (some) other featural groupings (e.g. ATR and nasality).These constraints are not intended as a direct adaptation of Padgett's account, but instead stand in as representative of the general idea of representations which have subgroupings.

For each randomly-generated word, up to seven candidates are generated. The faithful candidate is always included, as well as an instance of each type of harmony listed above. In addition, one candidate shows a CD harmony pattern: disharmonies of types C and D are both resolved. This candidate is important because it offers a contrast with the AB harmony candidate—the AB candidate (in a language that

Harmony system	Average residual error	Starting error
Harmony A only	0.242	0.308
Harmony B only	0.241	0.310
Harmony A and B	0.219	0.309
Harmony C only	0.329	0.308
Harmony D only	0.328	0.350
Harmony C and D	0.361	0.352
No harmony	0.401	0.465

Table 1.11: Average remaining errors after learning each of the toy harmony systems compared with error resulting from weight randomization. Weights generated as absolute value of standard normals.

prefers it) is supported by an additional constraint favoring that type of "conjoined" harmony, while the CD candidate (in a language that prefers it) is not. In many cases, harmony of one type or another is vacuous. These candidates are disregarded in the evaluation of results.

The learner is given 100 iterations at a learning rate of 0.01 to learn data generated in this way. For any given instance of learning, one of the types of candidates is chosen as the target for learning—that is, there are seven total language types considered, exhibiting each type of harmony and faithful mapping. The probability given to non-target forms is recorded. This is repeated 10,000 times for each of the seven systems considered. Results are given in Table 1.11.

As expected, Harmony A is just as learnable as Harmony B when these are learned alone, with residual error of approximately 0.241 on average. Similarly, Harmonies C and D each have error around 0.328 when they are the only harmony learned. When harmonies are learned in pairs, differences emerge. Harmony of both A and B, the analogue to color, is more learnable than harmony of C and D. The former has a residual error of only 0.219, while the latter remains at 0.361. It is clear that the distinct representational assumptions of the two types of harmony result in distinct learnability predictions. Encoding the "co-patterning" of A and B results in harmony

of a dual type being favored compared to a combination of two harmonies without a representation of co-patterning.

This pattern emerges particularly from learning, as can be seen by contrasting the distribution of learning results in Figure 1.7 with the distribution of results for initial weight settings in Figure 1.6. In initial weight settings, a difference is only seen between single harmonies, dual harmonies, and faithful systems. No distinction is made between types of dual harmony.

This simulation, though a simplification of the actual facts of harmony systems, is illustrative of how intuitive frequency explanations find a fuller interpretation in a learnability account. The apparent redundancy and grouping use of a color class is attractive for motivating phenomena which use this class. Learnability results verify that numerical predictions with this kind of basis are possible, even if they are typically omitted from discussion.

1.6.4 Dimensionality of representations

The discussion of harmony illuminates one way in which relative learnability results may differ. One type of learnability result is fixed with respect to a particular constraint set and merely compares how quickly learning proceeds on one selection of candidates compared with another. This is the type discussed in the stress learning results. The second type of learnability result instead compares learning speed across two constraint sets. In the case of harmony, this was the comparison of learning with and without a third HARMONIZE constraint referring to a shared class. This type of comparison, though not the major focus of this work, is important in its relation to typical theory building. Therefore, it is worth discussing the role of dimensionality—the number of constraints—in a learning problem.

If we assume that representations must fall within some bounded region of n-dimensional space, with equal bounds on each dimension, two points are maximally

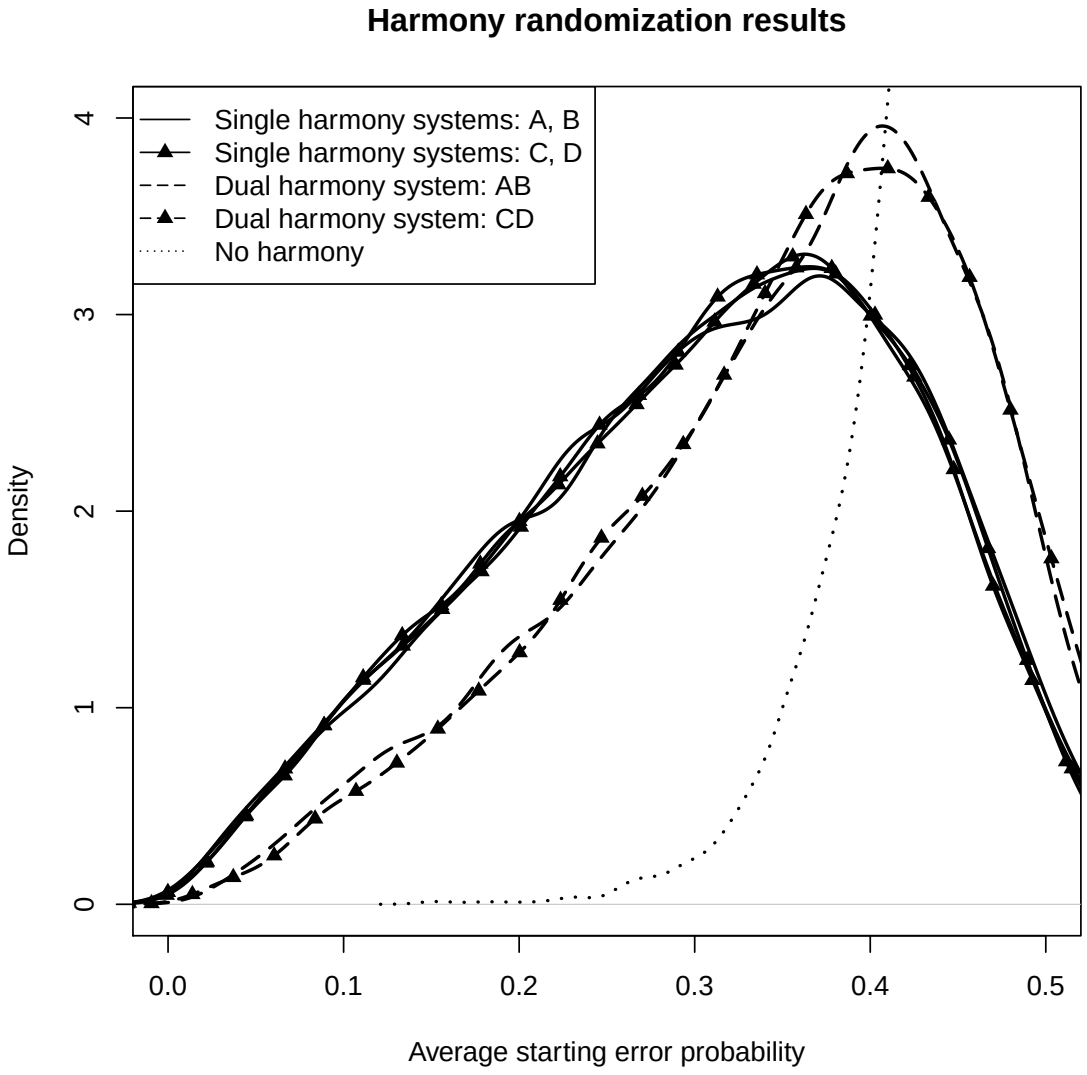


Figure 1.6: Distribution of average errors at starting weights for each of the toy harmony systems.

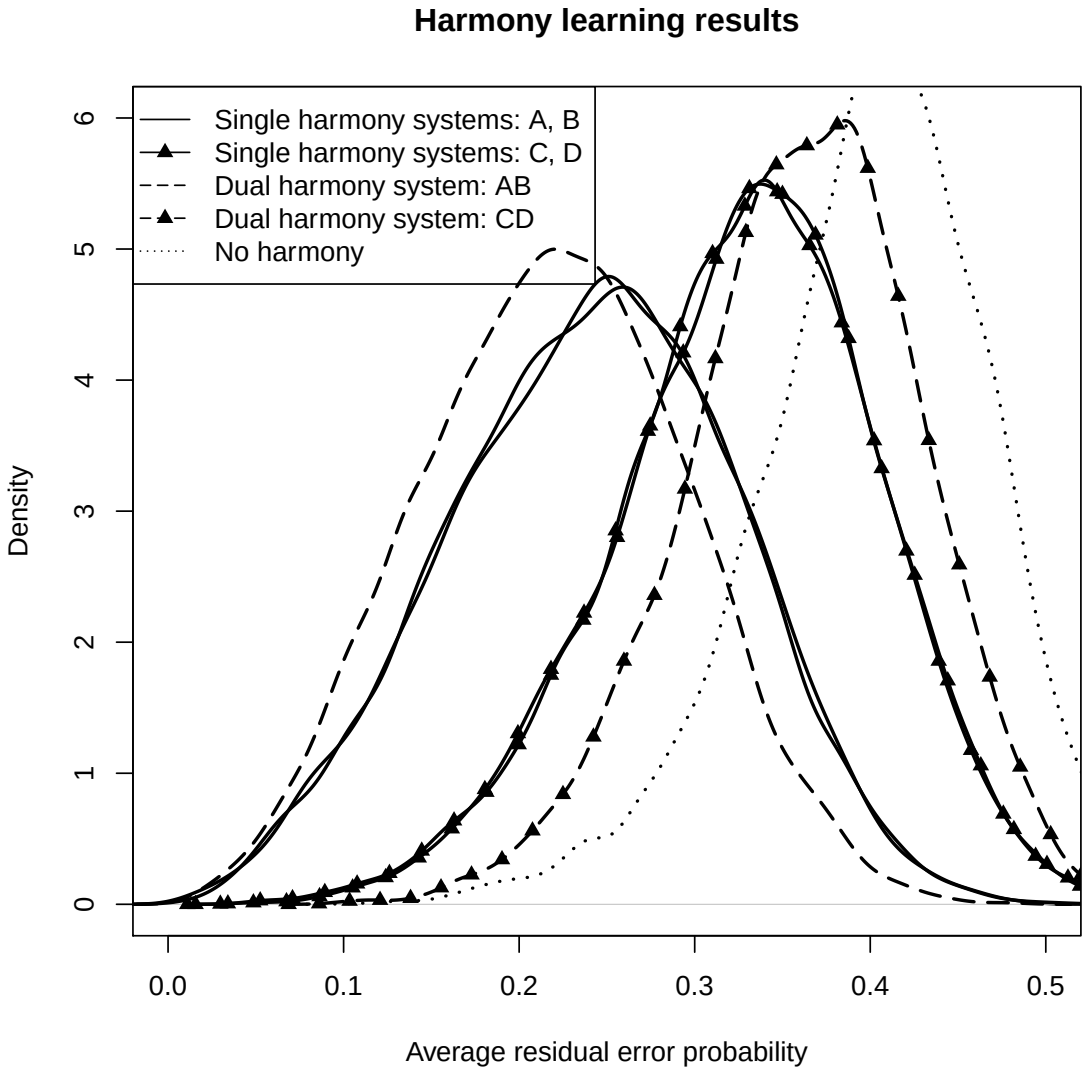


Figure 1.7: Distribution of average remaining errors after learning each of the toy harmony systems. $\eta = 0.01$.

distant if they are different by the full size of the bound on each dimension. For example, assuming a 3-dimensional unit cube bounded at 0 and 1 ($\mathbf{x} \in [0,1]^3$), we will have maximally distant representations like the pair $\langle 0,0,0 \rangle$ and $\langle 1,1,1 \rangle$. More generally, we can contrast one element with n zero values and another with n one values.

For the 1-dimensional case, the two points are obviously distance 1 apart. In two dimensions, this is a familiar case of the Pythagorean Theorem—the distance is $\sqrt{2}$. In general, the distance is just the square root of the number of dimensions.

$$\text{Distance} = \sqrt{\sum_{1}^{n} (1-0)^2} \tag{1.20}$$

$$= \sqrt{n} \tag{1.21}$$

The increased distance between points in higher dimensions has natural consequences also for probabilistic patterns. To see this, we can consider the *average* distance between two points in such a unit hypercube. First, let us assume the two points are chosen uniformly from the hypercube.

The majority of this expectation can be found using standard values for the expectation and variance of a $[0,1]$ uniform variate ($\frac{1}{2}$ and $\frac{1}{12}$, respectively).

$$\mathbf{X}, \mathbf{Y} \sim \text{Uniform}(0,1)^n \tag{1.22}$$

$$E\left[\|\mathbf{X}-\mathbf{Y}\|^2\right] = E\left[(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})\right] \tag{1.23}$$

$$= E\left[\sum_{i=1}^{n} X_i^2 - 2X_iY_i + Y_i^2\right] \tag{1.24}$$

$$= \sum_{i=1}^{n} E\left[X_i^2 - 2X_iY_i + Y_i^2\right] \tag{1.25}$$

$$= n\left(2E\left[X^2\right] - 2E\left[XY\right]\right) \tag{1.26}$$

$$E\left[X^2\right] = Var[X] + E[X]^2 \tag{1.27}$$

$$= \frac{1}{12} + \frac{1}{2}^2 \tag{1.28}$$

$$= \frac{1}{3} \tag{1.29}$$

It remains to find the expectation of the product XY. This is a product distribution and its density can be found using a method analogous to convolution, applied to a product.

$$X, Y \sim \text{Uniform}(0,1) \tag{1.30}$$

$$Z = XY \tag{1.31}$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx \tag{1.32}$$

$$= \int_{-\infty}^{\infty} f(x) f(z/x) \frac{1}{|x|} dx \qquad \text{identically distributed} \tag{1.33}$$

$$= \int_{z}^{1} \frac{1}{|x|} dx \qquad \text{pdf 1 iff } x \in [0,1] \tag{1.34}$$

$$= -\log z \tag{1.35}$$

We then find the expectation.

$$E[Z] = \int_{-\infty}^{\infty} z f_Z(z) dz \tag{1.36}$$

$$= \int_0^1 z f_Z(z) dz \tag{1.37}$$

$$= -\int_0^1 z \log z dz \tag{1.38}$$

$$= -\left(\frac{1}{4} z^2\right) (z \log z - 1) \,|_0^1 \tag{1.39}$$

$$= \frac{1}{4} \tag{1.40}$$

With all pieces in hand, we see that the average squared distance between two uniformly-distributed points in n-dimensional space is $\frac{n}{6}$. That is, squared distance increases linearly with the dimension of the space.

$$E\left[\|\mathbf{X} - \mathbf{Y}\|^2\right] = n\left(2E\left[X^2\right] - 2E\left[XY\right]\right) \tag{1.41}$$

$$= n\left(2\left(\frac{1}{3}\right) - 2\left(\frac{1}{4}\right)\right) \tag{1.42}$$

$$= \frac{n}{6} \tag{1.43}$$

From this presentation alone we do not know the expectation of the distance itself. Square root is a concave function, so by Jensen's inequality the distance is bounded above by $\sqrt{\frac{n}{6}}$.

If we consider normally-distributed points, the picture is much the same. With this distribution, the result emerges from standard identities relating normal, χ^2, and χ variates.

$$\mathbf{X}, \mathbf{Y} \sim \mathrm{N}(0,1)^n \tag{1.44}$$

$$\mathbf{X} - \mathbf{Y} \sim \mathrm{N}(0,2)^n \qquad \text{difference of two standard normals} \tag{1.45}$$

$$\frac{1}{2}\sum_{i=1}^{8}(X_i - Y_i)^2 \sim \chi^2(n) \qquad \text{sum of } n \text{ squared standard normals} \tag{1.46}$$

$$\sqrt{\frac{1}{2}\sum_{i=1}^{8}(X_i - Y_i)^2} \sim \chi(n) \tag{1.47}$$

$$E\left[\sum_{i=1}^{8}(X_i - Y_i)^2\right] = 2E\left[\frac{1}{2}\sum_{i=1}^{8}(X_i - Y_i)^2\right] \tag{1.48}$$

$$= 2n \qquad \text{mean of } \chi^2 \tag{1.49}$$

$$E\left[\sqrt{\sum_{i=1}^{8}(X_i - Y_i)^2}\right] = \sqrt{2}E\left[\sqrt{\frac{1}{2}\sum_{i=1}^{8}(X_i - Y_i)^2}\right] \tag{1.50}$$

$$= 2\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \qquad \text{mean of } \chi \tag{1.51}$$

$$\tag{1.52}$$

Thus for both uniform and normal distributions, the squared distance increases linearly with dimension—as $\frac{n}{6}$ for uniform, as n for normal—and the distance itself increases with the square root of dimension.

The increasing apparent size of spaces in more and more dimensions has implications for learning in the paradigm discussed. Increasing numbers of dimensions correspond to larger numbers of constraints. Each additional constraint gives the learner an additional potential manner in which to distinguish candidates. With greater and greater potential to distinguish the candidates, learning proceeds faster and faster. In terms of the learner's actual goal, this is clear. The learner seeks to find a hyperplane that divides the target language from the losers, maximizing probability. In higher dimensions, it is easier to "fit" a hyperplane between the two

classes. Additionally, due to the inflation of distance, many potential hyperplanes are essentially identical in terms of performance. With the space of possible solutions greatly expanded, the learner's task of finding valid weights can proceed quicker.

Another way to view the effect of dimensionality is at its limits. With no dimensions—that is, no constraints—we cannot distinguish candidates. Learning must therefore fail for any problem; it is "infinitely hard." In contrast, with an infinite number of constraints assigning violations to candidates in all possible patterns, learning is in some sense "infinitely easy." Every candidate will be uniquely preferred by *some* constraint, and every language will have infinite constraints in support of it.

This result is somewhat at odds with some intuitions about the task of the language learner. One view of linguistic theory is that it seeks to find the *necessarily very limited* mechanisms used in language representation because such limits are thought to improve learning. This view goes astray in two principal respects. First, it is not necessarily the case that limiting the search space improves the ease of learning, as discussed here. Such relationships between solution space and learnability are entirely due to assumptions about the learning algorithm—without explication of the learning procedure, such a goal is without a solid foundation. Second, it cannot be assumed *a priori* that—whatever factors *do* improve learning—human language is structured in such a way as to maximize learnability. Instead, we can only observe that language *is* learned and it is learned within some bounded amount of time. This cannot be strengthened to a claim that language is actually so as to be optimally learnable.

The work discussed here does not suggest the opposite view of theory building. It would not be well-motivated to increase the space of solutions for its own sake. Instead, it is useful to look for correspondence between a profusion of solutions and observed frequency in the typology. Indeed, as an *individual* language's number of possible representations increases, the proportional share of other languages must decrease. That is, to some extent making one language easier makes others harder.

1.6.5 Convergence and learning speed

I now elaborate the formal learning discussion begun in §1.6.2, making tighter connections with the perceptron convergence proof. I show that MaxEnt SGA is sufficiently like the perceptron learning algorithm that we can adopt much of the same mechanisms for showing approximate features characterizing its learning behavior. I explicate the perceptron convergence proof, showing where it fails for probabilistic grammatical theories like MaxEnt. I then show how analogous arguments can show the continued importance in MaxEnt for properties of the learning problem such as the size of the margin.

1.6.5.1 Lower bound

The goal of learning is to maximize the likelihood of the teacher's data under the learner's weights $\mathbf{w}$. However, in MaxEnt it is not strictly possible to maximize likelihood. There is no single maximum because any given arrangement of weights can be made more categorical simply by scaling: a set of weights like $\langle 1, 2, 3\rangle$ prefers the same candidates as $\langle 10, 20, 30\rangle$, only with less certainty.

Let us assume the existence of a $\mathbf{w}^*$ that comes within δ of reaching the maximum possible value of the likelihood (that is, 1). Further, let us assume that δ is chosen to be small enough such that all weight vectors reaching this criterion are parallel. Thus $\mathbf{w}^*$ performs well as a model of the data and all weights of similar performance prioritize individual constraints in the same way. The assumption of such a weight vector is justified for any learning problem that is well-modelled with a MaxEnt grammar and a given constraint set.

$$\exists \mathbf{w}^*, \delta : 1 - \prod_i p^*(i|\mathbf{w}^*) < \delta \tag{1.53}$$

$$\forall \mathbf{w}'. 1 - \prod_i p^*(i|\mathbf{w}') < \delta, (\mathbf{w}')^T w^* = \|\mathbf{w}'\|\|\mathbf{w}^*\| \tag{1.54}$$

To see that $\mathbf{w}^*$ is a sensible choice for an idealized goal of learning, consider how changing its weights affects performance on likelihood. We assume that the weight vector is close to fully describing the teacher's language. If all weights are multiplied by a common factor greater than 1, the harmony of target forms will increase and the harmony of non-targets will decrease. Thus, likelihood increases. This demonstrates that $\mathbf{w}^*$ is not uniquely defined. If the factor is less than 1, the distinction in harmony between targets and non-targets decreases. This reduces likelihood, potentially reducing it below $1 - \delta$. Finally, as all vectors that perform as well as $\mathbf{w}^*$ are parallel with it, any change to an individual weight will cause the vector to become non-parallel and therefore perform worse.

Based on $\mathbf{w}^*$ we can define a vector $\mathbf{u}$ which is a unit vector (i.e. of length 1) parallel with $\mathbf{w}^*$. For the minimum difference in harmony between a vector in the teacher's language and one outside it, using $\mathbf{u}$, write γ. This is the *margin*. For the maximum length of a violation vector considered, write ρ.

$$\gamma = \min_{i,j}(\mathbf{u}^T\mathbf{v}_i^* - \mathbf{u}^T\mathbf{v}_{ij}) \tag{1.55}$$

$$\rho = \min_{i,j}(\|\mathbf{v}_{ij}\|) \tag{1.56}$$

$$\tag{1.57}$$

This enables us to show a convergence result based on the perceptron convergence proof (Novikoff, 1962). First we see that the margin establishes a lower bound. For every iteration t at which there is an error (and thus, update), the weights can only change in a way bounded by the margin. The margin establishes (roughly) how far apart the closest vectors are. Any error will therefore have to make at least as much change to the weights as would be required by this closest pair.

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta(\mathbf{v}_i^* - \mathbf{v}_{ij}) \qquad \text{learning rule} \tag{1.58}$$

$$\mathbf{u}^T\mathbf{w}^{t+1} = \mathbf{u}^T\mathbf{w}^t + \eta(\mathbf{u}^T\mathbf{v}_i^* - \mathbf{u}^T\mathbf{v}_{ij}) \tag{1.59}$$

$$\mathbf{u}^T\mathbf{w}^{t+1} \geq \mathbf{u}^T\mathbf{w}^t + \eta\gamma \tag{1.60}$$

$$\mathbf{u}^T\mathbf{w}^{t+1} \geq \mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma \qquad \text{by induction} \tag{1.61}$$

Each additional error must at least increase $\mathbf{u}^T\mathbf{w}^{t+1}$ by γ. That is, every time the learner makes a mistake, it must increase the projection of its hypothesis onto a representation of the goal grammar by the margin. By induction we can thus eliminate a dependence on the previous step.

This result shows us that an increase in the margin γ increases the projection of $\mathbf{w}^{t+1}$ onto u. That is, an increase in the distance between candidates inside a language and the candidates outside it results in a reduced number of errors needed to achieve a weight vector parallel with the goal w^*.

1.6.5.2 Upper bound

In typical perceptron convergence proofs, an upper bound is also given in terms of ρ. This proof relies on the assumption that all errors were due to weights which produced an incorrect candidate, rather than random chance. It it thus not a true model of MaxEnt, but is nevertheless informative.

We start again with the update rule, but this time examine the sizes of the vectors.

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta(\mathbf{v}_i^* - \mathbf{v}_{ij}) \qquad \text{learning rule} \tag{1.62}$$

$$\|\mathbf{w}^{t+1}\|^2 = \|\mathbf{w}^t + \eta(\mathbf{v}_i^* - \mathbf{v}_{ij})\|^2 \tag{1.63}$$

$$\|\mathbf{w}^{t+1}\|^2 = \|\mathbf{w}^t\|^2 + 2\eta((\mathbf{w}^t)^T\mathbf{v}_i^* - (\mathbf{w}^t)^T\mathbf{v}_{ij}) + \eta^2\|\mathbf{v}_i^* - \mathbf{v}_{ij}\|^2 \tag{1.64}$$

The middle term of this equation is a difference of harmonies. The harmony of the correct form will always be lower, if we assume that an error occurred due to the weights rather than chance. Given this assumption, the very fact that an error occurred implies that the harmony of the non-target form is greater.

$$\|\mathbf{w}^{t+1}\|^2 = \|\mathbf{w}^t\|^2 + 2\eta((\mathbf{w}^t)^T\mathbf{v}_i^* - (\mathbf{w}^t)^T\mathbf{v}_{ij}) + \eta^2\|\mathbf{v}_i^* - \mathbf{v}_{ij}\|^2 \tag{1.65}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq \|\mathbf{w}^t\|^2 + \eta^2\|\mathbf{v}_i^* - \mathbf{v}_{ij}\|^2 \quad \text{by assumption} \tag{1.66}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq \|\mathbf{w}^t\|^2 + \eta^2\rho^2 \quad \text{definition of } \rho \tag{1.67}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 \quad \text{induction} \tag{1.68}$$

This bound can be combined with the previous one for a more accessible interpretation. We first state the margin result in terms of a norm:

$$\mathbf{u}^T\mathbf{w}^{t+1} \geq \mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma \tag{1.69}$$

$$\|\mathbf{u}\|\|\mathbf{w}^{t+1}\| \geq \mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma \quad \text{norm bound} \tag{1.70}$$

$$\|\mathbf{w}^{t+1}\| \geq \mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma \quad \text{unit vector} \tag{1.71}$$

Combining the two bounds gives us:

$$(\mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma)^2 \leq \|\mathbf{w}^{t+1}\|^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 \tag{1.72}$$

$$(\mathbf{u}^T\mathbf{w}^0 + (t+1)\eta\gamma)^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 \tag{1.73}$$

The importance of $\mathbf{w}^0$—the starting weights—vanishes as t grows large.

$$((t+1)\eta\gamma)^2 \leq (t+1)\eta^2\rho^2 \tag{1.74}$$

$$(t+1)^2\eta^2\gamma^2 \leq (t+1)\eta^2\rho^2 \tag{1.75}$$

$$t+1 \leq \frac{\rho^2}{\gamma^2} \tag{1.76}$$

Thus the time until convergence on $\mathbf{u}$ decreases with squared margin and increases with squared maximum violation vector size.

This proof can be modified for MaxEnt if we account for the probabilistic nature of output generation. The learner may be in error for two reasons: a legitimate issue with the weights or random chance. The former case falls under the above proof, while the latter can be accounted for if we bound the degree to which the optimization can stray from the Harmonic Grammar interpretation of a weight set.

This bounding can be simply accomplished if we switch to an alternate model of MaxEnt. In (1.3), the definition of MaxEnt probabilities is explicit—the exponentiated harmonies are simply normalized. Another interpretation for these probabilities are as the result of a latent variable model. In this interpretation, an optimum is chosen similarly to HG or Noisy HG. The learner computes a set of values G_{ij} from the harmonies, incorporating noise. It then chooses the greatest of these G_{ij} as optimal for the particular input i.

$$j = \underset{j}{\operatorname{argmax}}\, G_{ij} \text{ for a given } i \tag{1.77}$$

Unlike in Noisy HG, the noise ξ_{ij} is added to the harmony itself, not to the weights. The noise is independent and identically distributed for each candidate, not each constraint/optimization pair.

$$G_{ij} = H_{ij} + \xi_{ij} \tag{1.78}$$

$$= \mathbf{w}^T \mathbf{v}_{ij} + \xi_{ij} \tag{1.79}$$

To have the probability of each candidate match its MaxEnt probability, the noise ξ_{ij} should be standard Gumbel distributed (see e.g. Andersson and Ubøe, 2012), with the probability density function given in (1.81).

$$\xi_{ij} \sim \text{Gumbel}(0, 1) \tag{1.80}$$

$$f(\xi) = e^{-(\xi + e^{-\xi})} \tag{1.81}$$

We know that the learner's current form is preferred to the teacher's under some randomization. We can use the values G_{ij} to pull this apart.

$$G_i^* \leq G_{ij} \qquad \text{learner's chosen form won} \tag{1.82}$$

$$\mathbf{w}^T \mathbf{v}_i^* + \xi_i^* \leq \mathbf{w}^T \mathbf{v}_{ij} + \xi_{ij} \tag{1.83}$$

$$\mathbf{w}^T \mathbf{v}_i^* - \mathbf{w}^T \mathbf{v}_{ij} \leq \xi_{ij} - \xi_i^* \tag{1.84}$$

$$\tag{1.85}$$

The difference between two independent and identically distributed Gumbel variates follows the logistic distribution, so we can write this as:

$$\mathbf{w}^T \mathbf{v}_i^* - \mathbf{w}^T \mathbf{v}_{ij} \leq \psi_{ij} \tag{1.86}$$

$$\psi_{ij} \sim \text{Logistic}(0, 1) \tag{1.87}$$

This identity allows us to restart the second part of the proof from (1.65).

$$\|\mathbf{w}^{t+1}\|^2 = \|\mathbf{w}^t\|^2 + 2\eta((\mathbf{w}^t)^T\mathbf{v}_i^* - (\mathbf{w}^t)^T\mathbf{v}_{ij}) + \eta^2\|\mathbf{v}_i^* - \mathbf{v}_{ij}\|^2 \tag{1.88}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq \|\mathbf{w}^t\|^2 + 2\eta\psi^{t+1} + \eta^2\|\mathbf{v}_i^* - \mathbf{v}_{ij}\|^2 \quad \text{by (1.86)} \tag{1.89}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq \|\mathbf{w}^t\|^2 + 2\eta\psi^{t+1} + \eta^2\rho^2 \quad \text{definition of } \rho \tag{1.90}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 + 2\eta\sum_{k=1}^{t+1}\psi^k \quad \text{induction} \tag{1.91}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 + 2\eta(t+1)\frac{\sum_{k=1}^{t+1}\psi^k}{t+1} \tag{1.92}$$

$$\|\mathbf{w}^{t+1}\|^2 \leq (t+1)\eta^2\rho^2 + \|\mathbf{w}^0\|^2 + 2\eta(t+1)x^{t+1} \quad \text{CLT} \tag{1.93}$$

The last term involves the sample mean of independent random variables, so if t is sufficiently large its distribution is approximately normal by the central limit theorem. The mean of this normally distributed x^t is zero and its variance can be computed simply:

$$Var[x^t] = \frac{Var[\psi^k]}{t} \quad \text{Central Limit Theorem} \tag{1.94}$$

$$= \frac{\pi^2}{3t} \quad \text{logistic variance} \tag{1.95}$$

We can now return to (1.75) with the revised bound based on ρ.

$$(t+1)^2\eta^2\gamma^2 \leq (t+1)\eta^2\rho^2 + 2\eta(t+1)x^{t+1} \tag{1.96}$$

$$t+1 \leq \frac{\rho^2}{\gamma^2} + \frac{2x^{t+1}}{\eta} \tag{1.97}$$

Thus learning speed is characterized by the margin γ, the size of the largest vector ρ, the learning rate η, and random noise. We again see the term $\frac{\rho^2}{\gamma^2}$, specifically the

inverse relationship with the margin size. This validates a focus on the relative size of the margin in establishing the relative learnability of a particular language in this probabilistic setting.

1.7 Conclusion

In this chapter, I first introduced the problem of non-categorical typological data. Not all linguistic patterns are equally common, so a full model of linguistic typology needs to make predictions about frequency. The answer I propose is to model frequencies as emergent from the relative learnability of different languages under given assumptions about grammar. I explicated the particular model I adopt, a Maximum Entropy grammar learned online with SGA, as well as other concerns for modeling.

Following this, I introduced some typological tendencies of interest. These concerned simple patterns in the typology of stress—tendencies for fixed stress position, for certain types of alternation, and so on. I showed that a very simple set of n-gram constraints yields useful learning biases for explaining some of these tendencies. Such learners exhibit biases toward fixed stress near a word edge, perfect alternation, and more.

I developed an idea of why such learners should exhibit biases, setting out the most important feature of languages as their *distinctiveness*. A distinct language is one which is consistent in its pattern of constraint violation in a way that most other logically possible languages are *not* consistent. I show that this is not the only view of learning bias, giving an example of bias from redundant representations as applied to vowel harmony. Finally, I develop a more explicit formal understanding of the nature of the learning bias.

The choice of SGA in this work is not crucial. Any learning algorithm that produces some degree of variability in the results of learning should produce learning biases. Indeed, the biases discussed in this dissertation are not crucially linked to

SGA. SGA shows biases based on the relative distinctiveness of the target language with respect to its competitors, as discussed in §1.6.2. This bias relies on the fact that distinctive patterns will in general result in weight updates moving in consistent directions for the constraints which are important to an analyses. This is not a feature of SGA alone—SGA is in fact unlikely to be *most* influenced by such consistency. For example, multiplicative approaches like Winnow (Littlestone, 1988) disregard features which do not assist in learning, positively updating only the features that do. This means that—possibly more than SGA—the number of constraints with a consistent degree of violation will be very important to learning.

This chapter sets out the basic goals and methods of this dissertation. In later chapters, I will show other applications of modeled learnability to problems of probabilistic typology, particularly as applied to stress.

CHAPTER 2

EMERGENT TENDENCIES FROM GRAMMATICAL ASSUMPTIONS

2.1 Introduction

In this chapter, I consider the utility of particular grammatical assumptions for explaining tendencies in typology. These tendencies are shown to emerge from the use of grammatical representations by a learner, and require minimal additional stipulations. In a departure from the discussion in Chapter 1, the assumptions explored here are represented by a constraint set for stress motivated by typology, as typically discussed by analysts working in Optimality Theory. The constraint set represents some of the kinds of choices made by analysts faced by distinctions in the categorical typology, contrasting with the n-gram constraint set chosen from first principles.

The first bias considered pertains to methods of primary stress assignment. Primary stress is usually determined independent of reference to word parity (even or odd syllable count) or secondary stress location (e.g. van der Hulst, 1996). Instead of such reference, primary stress is usually placed on a designated privileged syllable: initial, final, penultimate, heavy, lexically marked, etc. I show that this bias in favor of so-called *non-counting* (e.g. Goedemans, 2010) systems emerges from learning if the grammatical assumptions include primary stress assigned on the basis of syllable-counting alignment (broadly construed). The bias emerges in learning because non-counting languages—represented in such a way—exhibit consistency, just as e.g. perfect grids exhibited consistency for n-gram constraints in Chapter 1. Counting languages will be inconsistent in their violation of primary stress alignment constraints, varying the position of primary stress depending on parity and word length.

Non-counting languages show less of this variability, with primary stress placed with respect to a consistent absolute position in the word. In addition to their relative consistency, non-counting languages have the additional property that most of their strings perform relatively well with respect to these primary stress alignment constraints. Taken together, this implies that the constraint-based representations of non-counting languages are relatively distinct from their alternatives, aiding learning.

The second bias considered relates to stress windows. The smaller a stress window is, the more common it is. Thus two-syllable windows are more common than three-syllable ones, and one-syllable windows (a degenerate case: initial and final stress) are quite common (Kager, 2012). These relative levels of attestation are accompanied by a seemingly categorical generalization: no windows of size four or above are attested. I show that the bias for small windows emerges from the concept of consistency: these windows are just those that cause different word lengths to be more similar. This probabilistic bias is sufficiently strong that the absence of windows of size four and above is unsurprising.

2.1.1 Probabilistic generalizations

Primary stress patterns can usually be described simply without reference to secondary stress. Primary stress tends to fall within a certain number of syllables of an edge, typically varying only due to properties of the syllables involved. It generally does *not* vary in placement based on word parity (whether word length is odd or even) or other facts related to secondary stress. That is, primary stress is largely independent of secondary stress and parity. StressTyp[1] (see Goedemans et al., 1996a; Goedemans, 2010) describes 85 iterative patterns as iterating from the left and 50 as iterating from the right. Of these, only 12 and 3 (respectively) require "counting" in

[1]Available (November 2013) at `http://www.unileiden.net/stresstyp/`.

primary stress that would motivate dependence on secondary stress or parity. This is a large asymmetry—135 languages versus only 15.[2] Such a pattern in typology compels linguistic analysis. If no languages had counting primary stress, a successful theory could simply exclude it. This is not the case, however; deviations are uncommon, but not absent. The probabilistic nature of the generalization does not mean it is not a generalization, however. It still calls out for explanation, and a complete theory of linguistic typology should provide one.

Another probabilistic generalization is found in the typology of stress windows. Two-syllable stress windows outnumber three-syllable ones, 121 to 39 in StressTyp. In this case, the existence of neither category is in doubt. Nevertheless, the asymmetry is profound enough that an explanation is clearly desirable. In addition to this probabilistic case, four-syllable windows (and larger) are categorically unobserved. One might ask, however, whether this gap is categorical by nature or an accident due to the low probability of such systems.

In this chapter I show that existing models of categorical generalization, combined with a general theory of grammatical learning, can address these sorts of questions. Such a model makes predictions about the frequency of linguistic patterns, allowing explanations of generalizations that are inexplicable in traditional typological modeling.

A theory designed for categorical predictions only gives a binary prediction of "possible" versus "impossible" for a given language type and can never make predictions about relative frequency over language types. Such models do not have such a goal and will not yield probabilistic predictions, despite any intuition to the contrary. In creating a linguistic theory, it is arbitrary to discard generalizations about frequency and probability, reducing them to questions of attestation. These typo-

[2]The number of counting systems drops to 11 when distinct varieties of Arabic are collapsed.

logical generalizations should not be ignored, even if they necessitate new ways of approaching data.

If a theory of probabilistic typologies is desirable, it must come from something supplemental to traditional categorical prediction models. As in Chapter 1, I propose that modeling probabilistic typology as the result of iterated grammar learning best accomplishes this expansion. The added complexity is comparatively minimal—the mechanics of categorical typology can be extended directly to be used in a learning system. In addition, this complexity is needed in *some* capacity in any case because we know that language is learned and passed down through generations of learners. Such a typological model simply incorporates independently-motivated components, implying that this view of probabilistic typology adds little or no added complexity to the system as a whole.

In the standard view of generative phonology, the task of the learning algorithm is to learn all and only the languages represented by a grammatical theory. In the approach demonstrated here, failed learning has value, pointing to distinctions in predicted frequency. This work thus follows suggestions such as that of Boersma (2003) that the explanation of typology would fruitfully be partitioned into the distinct responsibilities of a (fallible) learning algorithm and grammatical assumptions.

2.1.1.1 Accounts of probabilistic predictions

Though the generative tradition has focused primarily on issues of attestation, a concern with frequency is not new. Some part of the explication of frequency will inevitably be linked to historical and social circumstance. Linguistic theory should explain the mechanisms by which such forces are permitted to alter grammars, and also the range of grammatical variation they can create. However, it cannot necessarily be expected to answer why a particular set of contingent historical pressures came to be—this is just as true in the study of frequency as in the generative tradi-

tion more generally. In contrast, effects on frequency that arise from persistent facts about human biology or cognition do seemingly fall within the mandate of linguistic theory.

One source of explanation for typological tendencies in language is the nature of perception and production. These systems are constrained by the biology surrounding sound[3] and its related cognitive control. Taken together, this is the channel through which speech is transmitted. Properties of this channel could form direct substantive biases on learning or constitute a broader channel bias (Moreton, 2008) on typology. This sort of bias has been frequently implicated in language change, forming the basis of a theory of frequency in proposals such as Evolutionary Phonology (Blevins, 2004).

The work I present here is principally concerned with a different kind of learning bias: analytic or structural bias. This type of effect arises from the cognitive mechanisms associated with learning patterns, whether specific to language or not. These effects show up robustly in a number of artificial phonology experiments, particularly in relation to learning patterns described as the combination of features (Moreton and Pater, 2012). Iterated learning work (e.g. Kirby, 2002; Griffiths and Kalish, 2005; Kirby et al., 2007; Theisen et al., 2010), shows how biases imposed by transmission from teacher to learner (in theory and in experiment) can give rise to added structure over successive generations.

In recent years, a number of studies have pushed this type of explanation into phonological theory. Coetzee (2002) proposes that the typological frequency of patterns might be well-described by the relative number of rankings which describe a pattern, taken from an Optimality Theoretic factorial typology (Prince and Smolensky, 1993/2004). This line of investigation is taken up by Riggle (2008), formalizing the concept of "number of rankings" further into the notion of r-volume. Riggle

[3] *Sound* and associated terms stand in for any substance in which linguistic behavior is productively transmitted.

proposes a model of learning in which the r-volume of a pattern (its relative size in ranking space) determines whether it is chosen in cases of ambiguity. This parallels the evaluation metric of Chomsky and Halle (1968), which chooses shorter (or less informative) grammars in the face of ambiguous data. Moreton et al. (in prep.) advance a model of learning bias with a learning algorithm not specifically designed to account for bias.

The following modeling experiments demonstrate that even a quite general model of learning can account for substantial biases in stress typology. In this way, I follow Bane and Riggle (2008) in using stress as a lens on probabilistic typology due to its useful typological surveys and comparative abstraction from substance. However, I further claim that desirable typological predictions about frequency emerge from learning *in general*, not just models designed around their explanation.

2.1.2 Grammatical model

The model I adopt for learning simulations uses Maximum Entropy grammar (MaxEnt; Goldwater and Johnson, 2003). MaxEnt is a probabilistic version of Harmonic Grammar (HG; Legendre et al., 1990).

In the main simulations presented here, I use constraints chosen to represent standard sorts of distinctions in stress grammars, modifying the constraint sets of Alber (2005) and Kager (2005) specifically and McCarthy and Prince (1993a) and Prince and Smolensky (1993/2004) more generally.

1. Stress alignment: AlignFtLeft/Right: Assign a violation for every syllable between the left/right edge of a foot and the edge of the prosodic word.

2. Primary stress alignment: AlignHeadLeft/Right: Assign a violation for every syllable between the left/right edge of the head foot and the edge of the prosodic word.

3. Foot size: FtBin: Assign a violation for every monosyllabic foot.

4. Rhythmic: *Clash/*Lapse: Assign a violation for every pair of adjacent stressed/unstressed syllables.

5. Foot headedness: Iamb and Trochee: Assign a violation for every foot that is not strictly right/left headed.

I particularly call attention to the inclusion of Align constraints which count violations in syllables. This use of gradient alignment constraints conflicts with arguments to remove such constraints from a theory of Con (e.g. McCarthy, 2003, *pace).* However, this inclusion will aid in demonstrating the way that consistency can emerge and exhibit the kinds of biases desired—gradient Align is not necessarily the *only* way to do this.

The learning model used here is online—the learner processes each datum it receives in turn, adapting its hypothesis. A "teacher" randomly selects a word shape and produces a stress pattern for that word type based on its grammar. Shorter words are sampled exponentially more often than longer words, mirroring the distribution of word lengths in natural language. The learner considers a candidate set consisting of all metrical parses that include the correct number of syllables with one and only one primary stress. The learner produces its own parse for that word length. If the learner's predicted stress pattern does not match the teacher's, the learner updates its constraint weights. Learning is therefore error-driven: updates occur when expected and observed data do not match.

The learner's grammar is updated according to Stochastic Gradient Ascent (SGA; Jäger, 2007; Boersma and Pater, 2014), also known variously as the perceptron update rule or delta rule.

Some of the constraints used in the simulations involve foot structure, which is not overtly observable. This is problematic because the learner requires a foot structure to assign violations to the teacher's form. The learner must therefore make a decision about what hidden structure to use in evaluating the teacher's constraint violations. The approach presented here uses a probabilistic adaptation of Robust Interpretive Parsing (RIP; Tesar and Smolensky, 2000; Boersma, 2003; Jarosz, 2013; Boersma and Pater, 2014) to choose a likely hidden structure. In this version of RIP, the hidden structure used for a particular overt form is probabilistically chosen according to the grammar from all hidden structures consistent with the form. Thus the learner picks a foot structure compatible with the teacher's form that performs reasonably well according the learner's own grammar.

2.2 Bias from representation: primary stress and directionality

The most common source of categorical predictions for linguistic typology in the generative tradition is the nature of linguistic representations (e.g. Chomsky and Halle, 1968). Theories differ as to what form of representation is most relevant to typological prediction. In Optimality Theory, for example, typological prediction is performed by computing a factorial typology of the hypothesized constraints (Prince and Smolensky, 1993/2004). These constraints assign violations to representations, yielding particular patterns of violation—violation vectors—for particular candidates. These violation vectors are the lens through which OT views candidates and thus the "representation" relevant to looking at categorical predictions within such a theory. This type of prediction has an important role in probabilistic prediction as well (e.g. Riggle, 2008).

In learning-based typological prediction, representation has two distinct manifestations. First, representations operate as in standard theories in order to categorically

rule out certain patterns. Such results carry over simply to a learning framework: if a particular type of language is not representable in principle by the linguistic system, it will not be fully learnable as a categorical pattern. No amount of data will drive a learner to converge upon a hypothesis it cannot even entertain. That is, the typical means by which patterns are ruled "impossible" is still available. More distinctly, representation can affect the relative learnability of hypotheses. Depending on the structure of linguistic representation, a learner might gain more certainty for an "easy" hypothesis than a "difficult" one even with an identical amount of data.

I exemplify again this type of learning bias, first discussed in Chapter 1, using a strong typological tendency concerning primary stress. In general, the position of primary stress can be simply decided independently of the position of secondary stress or word parity. However, exceptions exist. Two languages exhibiting such patterns are Cairene Arabic (McCarthy, 1979) and Nyawaygi (Dixon, 1983a). These are languages in which primary stress must make reference to the same kind of iterative structure used by secondary stress. For example, in Nyawaygi secondary stress falls on every heavy syllable and in right-to-left trochees among light syllables. Primary stress falls on the initial or peninitial syllable, depending on whether it bears secondary stress. Thus primary stress is leftmost even while the general direction of parsing is right-to-left. This is in contrast to the dominant typological pattern linking the placement of primary stress with the direction of secondary stress.

This typological *tendency*—predominance of non-counting primary stress without the exclusion of counting—is impossible to account for in a theory which yields only categorical predictions for typology. I show that such a bias emerges from an online learning model and constraints which make reference to syllable-counting alignment. This result highlights an important aspect of such biases: predictions for probabilistic typology may differ based on hypothesized representations in just the same way as

Left	Right
$(\grave{\sigma}\sigma)\sigma$	$\sigma(\grave{\sigma}\sigma)$
$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$
$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$
$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$
$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$

Table 2.1: Contrast between trochees parsed from the left and from the right.

categorical typology. This fact offers a new tool for deciding theoretical questions based on typological data.

2.2.1 Typology of primary stress and directionality

Primary stress could conceivably be assigned on the basis of patterns of syllables, weight, etc. or on some combination of these properties and the placement of secondary stress. That is, primary stress could be regarded as independent of secondary stress or as (in some regard) parasitic on it. An argument for the first situation can be made on the basis of a tendency in the relationship between primary stress and directionality, when stated in terms of iterative foot parsing. Primary stress tends to fall on the "first" foot placed in a metrical parse—it does not depend on word parity (or length generally).

In iterative stress, the first foot is the place from which the system seems to count syllables for stress assignment. In Table 2.1, the two patterns are easily explained as alternating stressed and unstressed syllables (i.e. trochees) starting at the left or right edge, respectively. An attempt using the opposite edge yields a comparatively clumsy description: stress is penultimate or antepenultimate depending on word parity (and iterative thereafter). The avoidance of such reference to word parity as a concept in its own right is a motivation for both metrical and rhythmic approaches to stress (e.g. Liberman and Prince, 1977).

First foot/Non-counting	Last foot/Counting
$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$

Table 2.2: Contrast between primary stress on the first foot of a left-to-right parse and on the last one. Primary stress on the first foot requires no reference to syllable count (non-counting) while primary stress on last foot does (counting). Directionality is unimportant: the "first" foot of a right-to-left parse is the rightmost.

In the great majority of cases these systems place primary stress on the "first" foot of a parse. That is, a system which counts from the left will place primary stress on the leftmost foot (*vice versa* for the right). However, some exceptions exist, placing primary stress on the "last" foot parsed. This distinction is important for the question of whether primary stress can depend on syllable count. If primary stress is fully permitted to count syllables, with placement varying depending on word parity, there is seemingly no *a priori* reason to suppose that first foot languages should predominate. On the other hand, if primary stress is totally prohibited from counting, languages in which the last foot bears primary stress should be impossible.

Table 2.3 shows data from Apurinã (Facundes, 2000) and Wargamay (Dixon, 1983b), two languages sharing a secondary stress pattern in light syllables but differing in whether primary stress is counting. Apurinã reflects the dominant tendency for non-counting primary stress: primary stress does not vary with word parity in Apurinã, while it does in Wargamay.

Primary stress is rarely counting. In the StressTyp (Goedemans, 2010) database of stress patterns, only 15 of 120 iterative stress languages have counting primary stress.

These counts make it obvious that languages in which primary stress crucially depends on secondary stress or word parity are comparatively rare. In the non-counting

Apurinã (non-counting)		Wargamay (counting)	
ˈsɨto	‘woman’	ˈbada	‘dog’
paˈtaro	‘chicken’	gaˈgara	‘dilly bag’
ˌtakaˈtarɨ	‘manioc frying pan’	ˈgiɟaˌwulu	‘freshwater jewfish’
aˌnãpaˈnarɨ	‘dog’	baˈɟinɟiˌlaŋgu	‘spangled drongo-ERG/INSTR’
ˌnɨtaˌkapeˈrɨko	‘I will have put/planted it’	ˈjajimˌbaliˌlagu	‘play about-INTR.PURP’

Table 2.3: Apurinã (Facundes, 2000) and Wargamay (Dixon, 1983b) both stress every other syllable from the right (right-to-left trochees). Primary stress position varies with word parity in Wargamay but not in Apurinã.

	Iteration from left	Iteration from right
Non-counting	73	47
Counting	12	3

Table 2.4: Bias for non-counting iterative stress in StressTyp, divided by direction. No significant difference is claimed for direction.

languages, primary stress can be assigned without reference to this information. In such a pattern, primary stress is “independent” in the sense that it does not utilize other stress information. It is “non-counting” in the sense that no reference to parity is needed.

The frequency generalization about counting also extends to iterative bidirectional stress systems. Bidirectional systems are ones in which a foot is placed *opposite* the start of iteration. It is this foot which typically bears primary stress in a secondary stress system of this type. The other feet iterate toward this single “opposite-edge” foot. This sort of pattern is fully compatible with the idea of non-counting primary stress. The primary stress in the typical bidirectional systems can be determined without reference to secondary stress or parity—in fact, it can be considered as logically “before” secondary stress calculation.

Table 2.5 shows four schematic systems illustrating the differences between bidirectional and iterative stress assignment on the one hand and counting and non-counting primary stress on the other. It is clear that the position of primary stress does not

Non-counting primary stress		Counting primary stress	
Iterative	Bidirectional	Iterative	Bidirectional
$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	$(\grave{\sigma}\sigma)\sigma(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	$(\grave{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$

Table 2.5: Comparison between counting and non-counting primary stress for iterative and bidirectional systems. For standard iterative systems, non-counting primary stress falls near the edge at which iteration begins. For bidirectional stress it falls on the "stranded," opposite-edge foot.

depend on word parity in the non-counting primary stress languages: primary stress is always initial in these examples. This contrasts with counting primary stress, in which the location of primary stress is not simply expressible in terms of an edge. Instead, a primary stress description for these examples must make reference either to secondary stress or the length of a word.

The non-counting tendency can be expressed in several alternate formulations as a theory-internal generalization. In the tendency-obeying languages, primary stress can be placed without any reference to the placement of secondary stress—primary stress tends to be expressible in terms of a single privileged syllable (e.g. penultimate, initial, final heavy, etc.), rather than a particular privileged stress (e.g. the rightmost stress). In a derivational theory, the tendency means that primary stress usually falls on the first foot placed—giving an interpretation of the non-counting tendency as a "primary first" tendency. An alignment-based translation is more nuanced because left- and right-alignment are not the same as left-to-right and right-to-left parsing (Crowhurst and Hewitt, 1995; Alber, 2005). For example, a language with degenerate feet at the left edge may be better *aligned* with that edge but "parsed" right-to-left. This means that the generalization cannot always be precisely restated as for example "left-/right-

aligning secondary stress systems tend to have left-/right-aligning primary stress" (but see Gordon, 2002 on related generalizations, given a categorical treatment).

Van der Hulst (1996) proposes the "Primary First" theory of stress assignment, embracing the corresponding formalization (see also Pruitt, 2012, on discussion and intermediate formulations). Under this approach, primary stress is always the first stress assigned. With standard methods of assigning primary stress, this implies that primary stress in simple iterative systems must be at the start of iteration and bidirectional systems must place primary stress on the opposite edge—there would be no other options. This strict version of Primary First accounts for a wide range of the typological data on stress. The majority of stress systems do in fact place primary stress on their "first" foot, in a way compatible with non-counting primary stress. However, as the above counts and examples demonstrate, this tendency is not an absolute. To account for exceptions, accounts framed within the Primary First assumption are necessarily driven to ascribe independent reference to word parity to primary stress assignment. Avoiding this sort of provision is precisely the goal that motivated elaborated models of metrical structure and rhythm. Additionally, this undermines the explanation of the non-counting tendency: the theory is too permissive, allowing counting systems on an essentially equal footing with non-counting ones. Thus the probabilistic prediction approach is potentially very attractive: the crucially non-categorical bias for non-counting primary stress can in principle be accounted for without either ignoring counting cases or building redundant and under-supported theoretical mechanisms for primary stress assignment.

2.2.2 Results

Learning is as described in Chapter 1. The learning rate η is 0.1, as usual in this dissertation. Starting weights are drawn from a Gaussian distribution with mean

Non-vacuous	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$
Vacuous	$(\acute{\sigma}\sigma)\sigma\sigma\sigma$	$(\acute{\sigma}\sigma)\sigma\sigma\sigma$

Table 2.6: Contrast between vacuous and non-vacuous "flipping." Primary stress moves when put on the "opposite edge" of a trochaic parse, but does not when an initial stress pattern is flipped. In the latter case, there is no other foot to place primary stress on.

and standard deviation 10 truncated at zero (i.e. preventing negative weights).[4] The tableau for a particular word length consists of all parses of that word length using maximally binary feet.

A bias in learning for a particular linguistic feature, e.g. non-counting primary stress patterns, can be illuminated by comparing the learning of languages possessing and lacking that feature. For the purposes of the simulations here, I start with the quantity-insensitive stress patterns used by Bane and Riggle (2008), supplementing Heinz (2007). All of these languages are non-counting, so additional languages must be considered which might violate the generalization. In addition to these 26 stress patterns, I include the "flipped" version of 17 of them. These flipped patterns have identical secondary stress to some attested language but primary stress aligned toward the opposite edge. For example, Table 2.2 shows a language and its "flip"—both languages have trochees parsed left-to-right, but they differ in whether the leftmost or rightmost stress is primary. Apurinã and Wargamay in Table 2.3 give a real example of a flipped pair, when quantity sensitivity is ignored. The remaining 9 languages are not included in numeric results because they flip vacuously—for example, final stress remains final even if primary stress "moves" because there is only a single stress in the word to "move" to. This contrast is shown in Table 2.6.

[4]Other distributions were tested with few qualitative differences but I have made no full exploration of assumed weight distributions. This is likely to be a useful topic of inquiry.

	Mean Diff.	Diff. Range
Same-edge primary	−0.664	−1.429 to 0.932
Bidirectional	−1.145	−1.982 to −0.514

Table 2.7: Summary of non-counting bias results measured by SSE. A negative difference means the pattern in the typology was learned better. The flipped languages tend to be counting, suggesting bias.

This inclusion of flipping allows the comparison of languages differing only on their adherence to the non-counting tendency. Any difference in learnability must be attributed to differences in primary stress. Results are included in Figure 2.1. In this graph the residual error after some fixed amount of learning is compared between a language and its flipped counterpart (for example, the pair in Table 2.2 corresponds to a single dot). Error is reported as sum squared error (SSE). SSE is the sum over all stress patterns of the squared difference between the predicted probability of a stress pattern and its target probability. It thus summarizes the difference between two distributions in a single number: higher SSE indicates less successful learning of a pattern. The line in this graph represents a situation of equal error. This would be true of any pattern that was just as learnable as its flip. Significant deviation from the line indicates bias—one language or the other is better learned in the time allowed. A primary first bias is apparent here: the iterative languages are primarily above the line, indicating better learning in the unflipped, generalization-conforming pattern. I will return to the two exceptions in §2.2.3. These results are presented numerically in Table 2.7. In this table I present the difference between unflipped and flipped pattern errors, where negative values indicate non-counting bias. Heinz (2007) includes only generalization-conforming, non-counting systems, thus the (non-vacuously) flipped languages can be seen as approximating the set of counting patterns.

Another way to look at learning predictions is to examine the actual result of an attempted learning instance rather than error alone. That is, we may ask what languages are likely to be learned when given certain kinds of input. In the simplest

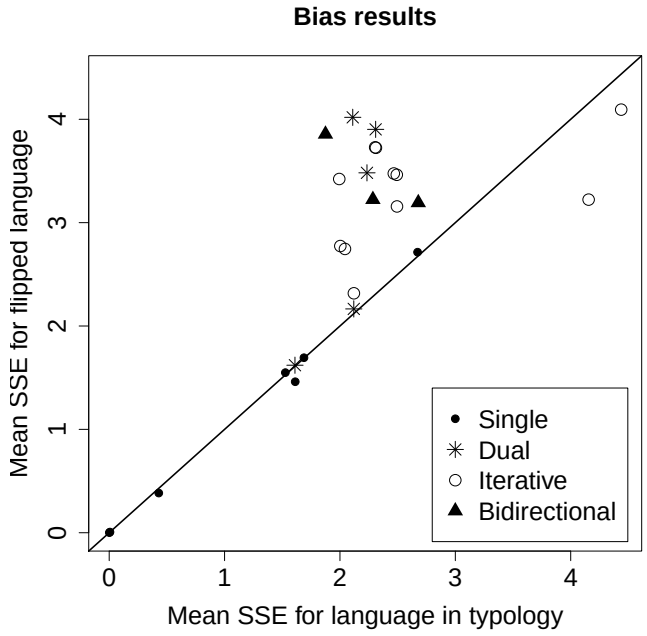


Figure 2.1: Learning bias in favor of generalization-conforming languages. Points above the line show bias in favor of the language in the typology; points below show bias in favor of the "flipped" language. Single stress languages have one stress per word, dual have at most two, iterative & bidirectional have stresses in proportion to word length. Single stress languages are included to estimate noise. $\eta = 0.1$, 1,000 trials, 1,000 iterations each.

case, all languages are learned faithfully: given data from a particular language, a learner would only produce a grammar exactly consistent with the data provided by its teacher. The degree to which individual languages differ from this ideal is their learning error, discussed above. We may also look at what languages are likely to be produced from particular sources other than the original language. That is, we ask whether a particular pattern is likely as a failed instance of learning another pattern. Rafferty et al. (2011) show that in iterated learning models of typology this is a necessary step: simple error is not sufficient to describe typological trends. This is the model of iterated learning described in Chapter 1.

Figure 2.2 shows results of this kind. This figure shows the probability of acquiring a particular language given a particular language as the data source, giving this prob-

ability as shading. Learning is carried out with a teacher for each of the languages in the augmented typology described above. The resulting grammar is then compared with the languages of the typology. The language which receives the maximum likelihood[5] under the learned grammar is counted as the resulting language. The figure presents a confusion matrix of these results: the probability that a particular initial language yields a learned grammar that best describes some particular (potentially different) language. The columns represent particular starting languages, the rows particular resulting language, and the darkness of the square the frequency of a result. The diagonal of this matrix represents faithful learning. The upper-right box represents "flipped" languages learned as flipped languages. The rectangle below it shows flipped languages learned as unflipped languages and the one to the left shows unflipped languages learned as flipped. Of note is the fact that the former unfaithful box is comparatively filled and the latter is comparatively empty. That is, mislearning is more likely to produce generalization-conforming (non-counting) languages than non-conforming (counting) ones.

It is clear that these flipped languages are not completely impossible to learn in a single generation. This is as expected given the gradient nature of error in Figure 2.1. To further probe predictions of typology, iterated learning can be modeled. In such a model a learner's final grammar is used to generate data for a second generation learner, which in turn generates data for a third, and so on. Figure 2.3 shows results for this kind of learning model. A concentration of probability onto fewer languages is apparent, with a move particularly away from languages which contradict the non-counting tendency. Graphically this presents as less faithful learning (on the diagonal) and some movement away from the boxed tendency-disobeying languages.

[5]Here the maximum likelihood criterion is used, rather than exact string equality, because resulting grammars may differ in small ways not relevant to the typological comparison in question.

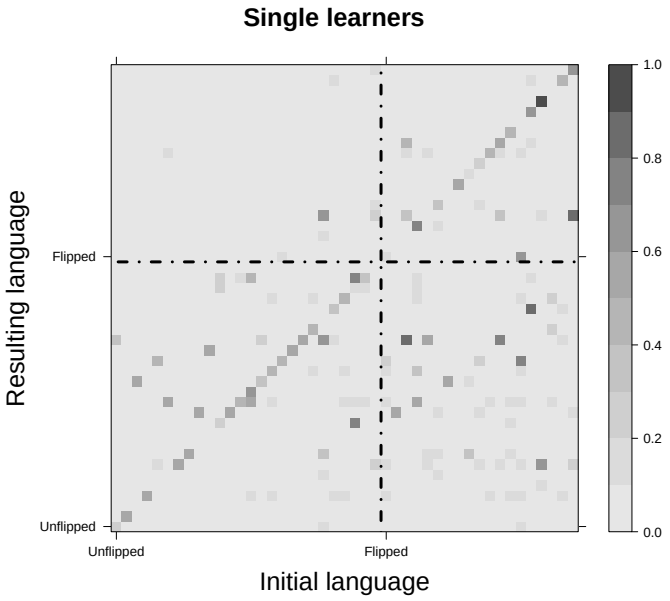


Figure 2.2: Single step learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language. 500 trials with 10,000 iterations per trial. Lower-left: unflipped language learned as unflipped. Upper-right: flipped language learned as flipped. Upper-left: unflipped language learned as flipped. Lower-right: flipped language learned as unflipped. $\eta = 0.1$.

A final way of looking at this sort of bias systematically is to calculate the theoretical results of iterated learning based on a single instance of learning. To do this, we take the matrix in Figure 2.2 as representing the probability that a learner categorically learns a language based on some initial language. We can then calculate the probability of future generations acquiring each language by exponentiating this matrix (see Chapter 1). This method has the advantage of easily calculating long-term predictions. A principal disadvantage is that this use of learning probabilities assumes that every language under consideration is necessarily learned as one of the (possibly distinct) languages in the set and that the learned grammar is categorical. However, results for this problem are qualitatively similar to simulated

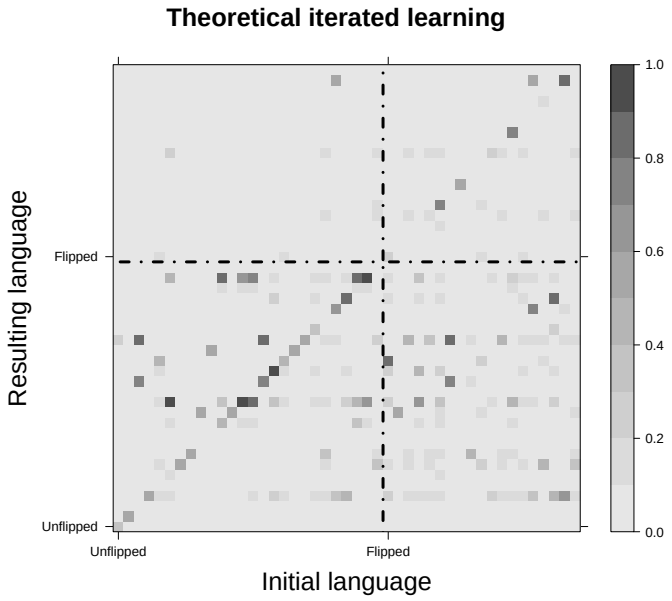


Figure 2.3: Simulated iterated learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language estimated over 100 trials. "Flipped" languages are boxed. $\eta = 0.1$, 10 generations with 10,000 iterations each. Lower-left: unflipped language learned as unflipped. Upper-right: flipped language learned as flipped. Upper-left: unflipped language learned as flipped. Lower-right: flipped language learned as unflipped.

iterated learning without this assumption.[6] With this tool in hand, we can calculate the probability of arriving at an iterative language obeying the primary first tendency compared with one disobeying it. Figure 2.4 shows these theoretical probabilities over many generations assuming a uniform initial distribution over languages.[7] Despite an equal number of flipped and unflipped iterative stress languages, the probability distributed over the unflipped languages is greater. Over time the flipped systems

[6]Iterated learning simulations were carried out, using the state of the learner at the end of one instance of learning as the teacher's state for the next. No substantive results claimed here differed for these simulations.

[7]Results are qualitatively similar with other starting distributions such as restricting the starting languages to iterative stress and/or using the frequencies of the unflipped languages from Heinz (2007) (the Stress Pattern Database).

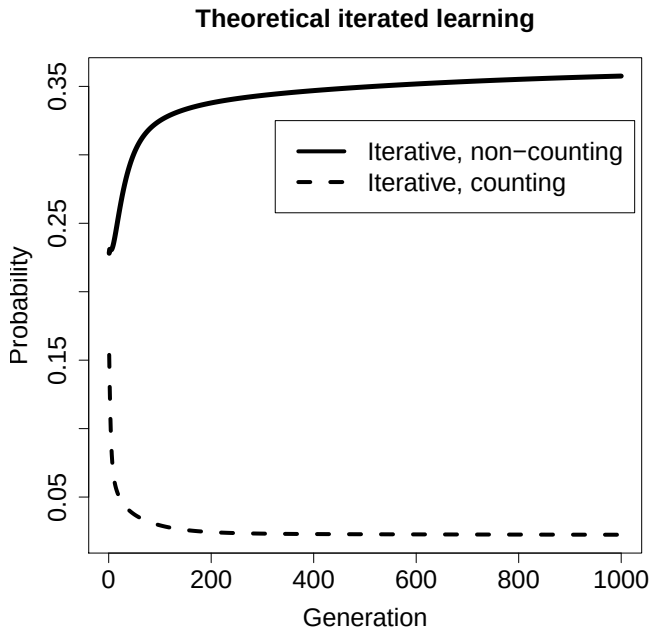


Figure 2.4: Theoretical bias of iterated learning. Probability distributed over all tendency-conforming iterative languages compared with all tendency disobeying ones. Calculated from learning results of Figure 2.2. The lines track the probability of iterative languages contrasted between the top and bottom of those graphs. $\eta = 0.1$.

yield probability to the unflipped ones. The frequency of the unflipped patterns does not increase arbitrarily because they compete with non-iterative systems.

2.2.3 Explaining the bias

In the above sections I have shown that the non-counting bias emerges in learning stress patterns. The origin of this bias can be better understood through consideration of violation vectors, the lists of constraint violations characterizing each candidate. A classic result by Novikoff (1962) gives a convergence guarantee for perceptron learning when applied to linearly-separable classes. This result shows a link between the speed of perceptron learning and properties of the vectors considered for classification. Learning speed decreases when the norm (roughly, the size) of the largest vector under consideration increases. Speed increases when the margin between vectors in one class and those in another class increases. this explanation is elaborated in Chapter 1.

These two considerations apply simply to gradual learning of constraint-based grammars. The size of the largest vector considered corresponds roughly to the most-violating candidate that must be considered in learning. That is, the wider the range of candidates that need to be ruled out or accepted, the slower that learning proceeds. The margin between vectors corresponds to the distinctiveness of the candidates which are part of a language from those which are excluded from it. Thus the speed of learning increases when candidates in a language are all similar to one another (in terms of violations vectors) in ways in which they are dissimilar to candidates out of the language. A margin of separation of this type has a role in the optimization work of Potts et al. (2010) and the adapted perceptron convergence proof for Noisy HG of Boersma and Pater (2014).

The important criterion for the non-counting tendency is distinctiveness. The candidates in a tendency-obeying language are generally distinct from the candidates outside of it. The reason is that primary first languages are more consistent across word length in their placement of primary stress. The first foot in a directional parse will be aligned closely with a word edge in a similar or identical way across different word lengths. If primary stress falls on this foot, constraints referring to primary stress will be consistently violated (or unviolated) in a way distinct from candidates outside the language. The strings within a tendency-obeying language are more consistent than those outside of it: they are distinctive. This distinctiveness produces the learning bias presented here. The non-counting bias in learning is thus closely related to an observation by Gordon (2002) that stress patterns placing stress a uniform distance from the word edge are preferred typologically.

The relationship of this idea of consistency or distinctiveness can be made more clear by considering two cases: bidirectional stress and primary stress clash. These two patterns are presented schematically in Table 2.9. Bidirectional stress is the chief reason for proposing that primary stress should actually precede secondary stress in a

		AlignHeadLeft	AlignHeadRight
Non-counting	$(\acute{\sigma}\sigma)$	**0**	1
	$(\acute{\sigma}\sigma)\sigma$	**0**	2
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	**0**	3
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	**0**	4
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	**0**	5
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	**0**	6
Counting	$(\acute{\sigma}\sigma)$	0	1
	$(\acute{\sigma}\sigma)\sigma$	0	2
	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	2	1
	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	2	2
	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	4	1
	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	4	2

Table 2.8: Comparison of violation vectors for counting and non-counting versions of a single secondary stress pattern. Strings from Table 2.2. The non-counting language has a consistent pattern of violation for AlignHeadLeft, while the counting language has no corresponding consistent constraint. All other constraints are omitted because their violations are exactly the same between the two patterns.

derivational account. The most common sort of bidirectional stress language is like the one presented—primary stress falls on the "stranded" foot towards which secondary stress iterates. This aligns with learning results in Figure 2.1—bidirectional stress is better learned in its more attested form. This follows from consistency of primary stress placement: such bidirectional systems have extremely uniform primary stress, while their flipped patterns can vary in primary stress placement.

The other pattern here shows the type of language in which a single instance of learning predicts a reversal of the primary first tendency—the two languages noted in §2.2.2 as being on the "counting-preferring" side of the line in Figure 2.1. One such language is schematized in Table 2.9. This language could be described as right-to-left trochees tolerating degenerate feet, but primary stress is on the "last" foot (i.e. the leftmost foot, as the pattern is right-to-left). This language—and its syllable-wise reversal, left-to-right iambs with degenerate feet and primary stress on the rightmost foot—are better-learned than their unflipped counterparts with stress on the "first"

Bidirectional		Iterative trochees	
Non-counting	Counting	Reduced $\acute{\sigma}$ clash	Increased $\acute{\sigma}$ clash
$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$
$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$	$(\grave{\boldsymbol{\sigma}})(\acute{\boldsymbol{\sigma}}\sigma)$	$(\acute{\boldsymbol{\sigma}})(\grave{\boldsymbol{\sigma}}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$
$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)\sigma(\acute{\sigma}\sigma)$	$(\grave{\sigma})(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\acute{\boldsymbol{\sigma}})(\grave{\boldsymbol{\sigma}}\sigma)(\grave{\sigma}\sigma)$
$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$
$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\grave{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\grave{\sigma})(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)(\acute{\sigma}\sigma)$	$(\acute{\boldsymbol{\sigma}})(\grave{\boldsymbol{\sigma}}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$

Table 2.9: Two important cases for the non-counting bias. Bidirectional stress is best learned in the better-attested non-counting form, with the primary stress foot consistently placed. For the iterative case, the system with increased primary stress clash (bolded) is learned better, but is less attested. Both iterative systems are non-counting.

(i.e. rightmost) foot. These patterns have clash between a primary stress and a secondary stress. As noted, these languages have increased primary/secondary stress clash compared with their unflipped alternative. Such languages are uncommon, as noted by e.g. Kager (2001). They are potentially perceptually problematic—the two stresses must be correctly perceived as two stresses of different types. However, they are very consistent: primary stress always falls in exactly the same position with respect to the word edge. In fact, these languages are not really counting. No reference is needed to word length in order to assign primary stress. As such, the nature of the bias in learning is unaltered. These languages are very rare but attested, for example South Conchucos Quechua (Hintz, 2006). Here perceptual and formal biases are in tension: the existence of such perceptually dispreferred languages may be due to their advantages in learning.

As a final point on this bias, it is important to realize that the constraint set is crucial to explaining learning results. The results here largely do not depend on the *particular* constraint set assumed and hold across a variety of assumptions. However, that does not mean that *all* constraint sets exhibit a bias toward non-counting primary stress. The constraint set must allow the learner to be biased by the regularity of

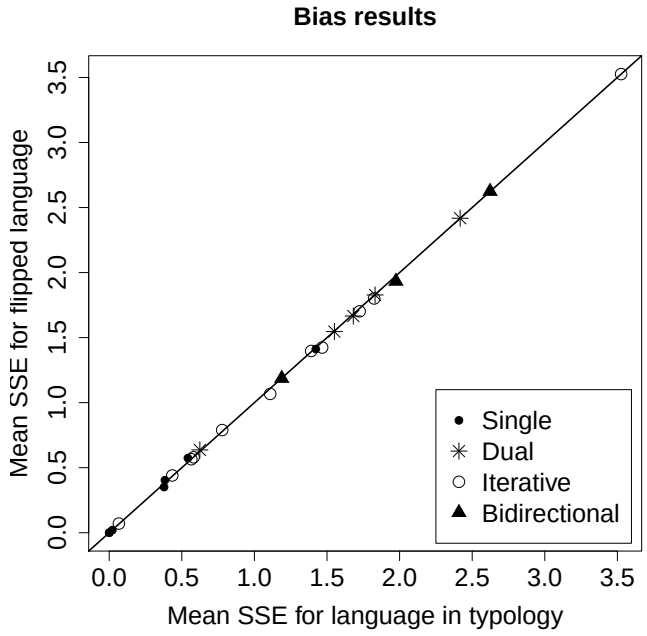


Figure 2.5: Learning bias in with Gordon constraint set. Points above the line show bias in favor of the language in the typology; points below show bias in favor of the "flipped" language. $\eta = 0.1$, 100 trials, 1,000 iterations each.

stress with respect to the word edge. The constraints proposed by Gordon (2002) for quantity-insensitive stress *do not* include a constraint that controls the position of primary stress relative to an edge. Instead, primary stress is placed according to an end rule (Prince, 1983) constraint over secondary stresses.[8] This leaves this constraint set without a way in which to represent consistency in primary stress placement, meaning that the bias is not predicted for such constraints. This result is demonstrated in Figure 2.5, corresponding otherwise exactly with Figure 2.1. The error of this learner does not distinguish flipped languages from unflipped ones. Thus with such a constraint set no account of this bias can be made with learning.

[8]Gordon limits his factorial typology to rankings with consistent use of directionality across constraints. I include both left- and right-oriented constraints in the simulations described. This is a possibility admitted by Gordon and it is more suitable to the learning algorithm used. However, this prevents the resulting typological model from echoing some of Gordon's target categorical generalizations.

	σσσσσ̲	σσσσ̲σ	σσσ̲σσ	σσ̲σσσ	σ̲σσσσ
"One-syllable"	σσσσσ́	σσσσσ́	σσσσσ́	σσσσσ́	σσσσσ́
Two-syllable	σσσσσ́	σσσσ́σ	σσσσσ́	σσσσσ́	σσσσσ́
Three-syllable	σσσσσ́	σσσσ́σ	σσσ́σσ	σσσσσ́	σσσσσ́
Four-syllable	σσσσσ́	σσσσ́σ	σσσ́σσ	σσ́σσσ	σσσσσ́

Table 2.10: Examples of window stress systems. If the designated property (underline) is within the window it is matched by surface stress. Otherwise default stress results. The default assumed here and throughout is edgemost.

2.3 Frequency and gaps: stress window size

In the preceding section I consider only attested quantity-insensitive stress systems and their flipped counterparts. These are not the only sort of stress systems which can benefit from a probabilistic, learning-based approach. In this section I consider stress window systems: languages in which stress is required to fall within a given number of syllables from an edge, but in which the choice of *which* particular syllable is made based on some other property of the syllables or word. This property might be quantity (weight), sonority, or lexically marked stress. In general I will refer to this property as a designated property, without theoretical commitment to its representational interpretation.

In my discussion of stress windows, I exclude a type of language exemplified by Axininca. Axininca (Payne, 1990; Hayes, 1995) main stress exhibits something very close to the kind of windows discussed. In this language, feet are left-to-right iambic and stress is nonfinal. Main stress is placed on one of the last two feet, whichever is heavier. This means that there is maximally a four-syllable window at the word edge in which stress can occur, determined by weight (Table 2.11). However, the manner of stress assignment is distinct from more general windows: the full window effect is only obtained when other factors allow. Instead of a four-syllable window, Axininca could perhaps be better thought of as a two-*foot* window. These types of cases are not included in the typological counts for stress window systems.

$$\begin{array}{lcl} ...\sigma\underline{\sigma}\sigma\sigma\sigma & \rightarrow & ...(\sigma\acute{\sigma})(\sigma\grave{\sigma})\sigma \\ ...\sigma\sigma\sigma\underline{\sigma}\sigma & \rightarrow & ...(\sigma\grave{\sigma})(\sigma\acute{\sigma})\sigma \end{array}$$

Table 2.11: Schematic view of Axininca main stress. Choice between final or penultimate foot, combined with nonfinality, creates apparent "four-syllable window."

Window type	Count	
Final two syllables	82	e.g. Yapese (Jensen et al., 1977)
Final three syllables	38	e.g. Pirahã (Everett and Everett, 1984)
Initial two syllables	39	e.g. Malayalam (Asher and Kumari, 1997)
Initial three syllables	1	e.g. Comanche (Smalley, 1953)

Table 2.12: Typological counts for window stress from StressTyp. Adapted from Kager (2012, ex. 22). Counts are collapsed across types of designated property and the position of default stress.

Stress windows show several marked asymmetries. The generalization of most theoretical relevance is the maximum size of windows. Languages which constrain stress sensitive to a designated property to a single syllable from the edge are extensionally equivalent to the amply-attested initial and final stress. Larger two- and three-syllable windows are well known. Despite this, no four-syllable window systems are known. Any full account of stress typology should deal with this absence.

Within attested windows, two facts are apparent. First, for a given size, windows are better-attested on the right edge than the left. This generalization perhaps echoes the dominance of penultimate stress over peninitial stress (see e.g. Gordon, 2002). Second, for a given word edge two-syllable windows are better-attested than three-syllable ones.

The relationship between window size and frequency is not surprising given the statistics of learning. As the size of a window increases, larger and larger words are needed to detect it. For example, a four-syllable window requires words of five and more syllables before it is distinguishable from a truly unbounded stress sensitive to designated properties. This fact alone limits possibilities for large windows—natural

linguistic data is not rich in long words in most languages, restricting the possible range of patterns even in principle (Hammond, 1991).[9]

As noted by Prince (1993, p. 12), the learning problem in a weighted grammatical model also makes larger windows more difficult to acquire. As window length increases, a narrower and narrower range of weights describes a pattern. Analogously, the mutual reliability of stress data degrades as window size increases. With short lengths, most strings surface as some default pattern, leaving only a small excluded class to be learned semi-independently. With large window size, in contrast, much surface data will be fully specified by a designated property. In such a situation, the learner receives data which does not support a particular window size, resulting in slower learning. This is analogous to the discussion of distinguishability in Chapter 1. The learner has two subpatterns to learn—default stress and sensitivity to a designated property. Learning is difficult when these two subpatterns are numerically equivocal.

Under a learning-based account of the frequency of window sizes, the absence of four-syllable windows (and larger) is potentially surprising. Legendre et al. (2006) point out that a simple Harmonic Grammar system with alignment constraints can model stress windows of arbitrary length, arguing that this sort of prediction poses a problem for the use of HG in typological prediction. Pater (2014), following Prince, contends that the learning problems posed by larger windows could explain the disconnect, but this is not a complete explanation. Even if four-syllable windows are unlikely, they are not necessarily impossible within an HG-like system that incorporates learning. It is therefore necessary to understand why these systems which are predicted to be only uncommon are completely absent. I show that this absence

[9]Short words tend to be much more frequent than long words, offering a tentative approach the frequency of short windows. However, in additional simulations the learning bias attributable to frequency was found to be numerically dominated by the reliability bias approach presented here.

is plausibly an accidental gap, given simulated results for iterated learning. Importantly, these languages are learnable to some degree, but their iterated transmission across generations is unstable. This solution to learning is not entirely unique: Pater (2009, fn. 9), for example, suggests a fixed margin required of all successful grammars. My account has the advantage of explicitly modeling theorized connections between learning and typology.

Such learning results are valuable because the (ostensibly) maximal three-syllable window poses difficulty for both HG and OT. In HG the problem is overprediction—it can model windows that are larger than those observed in languages of the world. However, OT has a problem of *under*prediction—the three-syllable window itself is not readily generated in typical OT constraint sets (Kager, 2012). Just as in §2.2, incorporating learning into typological prediction can improve a model's fit to the observed counts of language types.

2.3.1 Stress window simulations

Simulations showing bias in stress windows follow roughly the same form as those in the previous section. A MaxEnt learner is repeatedly exposed to data from the typology and its resulting error or final language is recorded. This gives a measure of bias on both single-step and typological bases. The foot-based alignment constraint set in §2.1.2 is used again.

The languages of interest are those which most simply demonstrate window size. These languages place stress on a syllable bearing a designated property within a window and on the edge otherwise. This is not necessarily the most common default (in fact, this is not likely). In these simulations, the learners only receive strings with a single designated property or with no designated property. In addition, the relevant effects are observable with just a single stress, so only such candidates are included. These simplifications mean that the full range of window stress patterns, especially

including variation in the frequency of patterns of default stress in window systems (Kager, 2012), is not modeled. However, this means that relevant languages are easily comparable.

The typology used for testing includes: fixed stress one to eight syllables from the edge (left or right) and window stress two to eight syllables from the edge (left or right). In the figures below, the eight left-counting fixed stress languages are followed by eight right-counting fixed stress, then the left and right windows. Within these groupings, the relevant syllable count increases moving rightward.

Figure 2.6 shows bias in learning outcomes. Fixed systems of count one and two (initial, final, peninitial, and penultimate) are learned faithfully. Larger fixed systems are mislearned as opposite-edge stress—e.g. pre-antepenultimate stress is mislearned as initial. These fixed systems do not interact with the window systems—window stress is not mislearned as fixed, and vice versa. Window stress shows more diffuse learning—higher syllable-count languages are learned unfaithfully, but are not consistently learned as one thing or another.

The directionality of these biases can be exposed, as in the previous section, by exponentiating the transition probability matrix. Figure 2.7 shows the 100th power of Figure 2.6. The near-categorical outcomes of fixed stress do not change considerably. Stress windows, however, consolidate probability toward lower syllable counts. Here two-syllable windows are learned faithfully, while larger windows are learned predominantly as three- or four-syllable windows.

This consolidation effect increases across successive generations. Figure 2.8 shows the relative probability of each window length taken over the window stress languages in general. Direction can be safely ignored here due to the symmetry of the constraint set in alignment. Windows larger than size three rapidly lose their share of the total probability. In this simulation, three-syllable windows gain an early boost to probability, eventually losing to two-syllable windows in the long run.

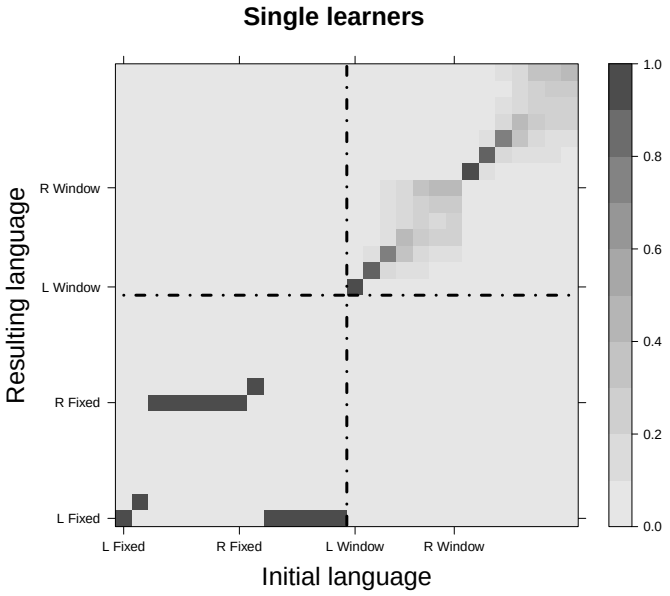


Figure 2.6: Single step learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language. Within each labeled category, number of syllables in the window increases to the right. $\eta = 0.1$, 2,500 trials per language, 2,500 iterations each.

The early rise of three-syllable windows is due to their role as a transitional state between larger windows and smaller ones. Windows of length four to eight will necessarily pass through three-syllable windows, even if ultimately arriving at a two-syllable state. These large windows are given equal probability to small sizes, inflating the three-syllable probability. The added level of scrutiny made possible by this kind of size-based generalization shows the bias generated by the (probably unreasonable) uniform starting distribution. In any case, the long-term dynamics are as expected. These effects do not depend on the starting distribution.[10]

[10] An iterated learning simulation starting from random strings predicts similar effects without an initial bias toward three-syllable windows. This is the methodology employed by e.g. Theisen et al. (2010) for experimental iterated learning.

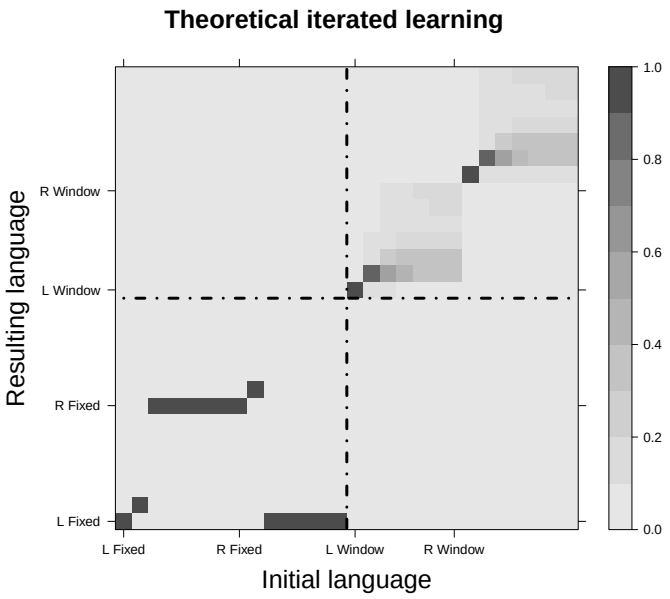


Figure 2.7: Simulated iterated learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language estimated over 100 trials. Within each labeled category, number of syllables in the window increases to the right. $\eta = 0.1$, 2,500 trials per language, 2,500 iterations each.

2.3.2 Explaining typological tendencies

These simulations give credence to the typological explanations proposed above. Short windows are more common typologically and are predicted to be so by such a model of learning. These types of windows give greater reliability on learning data across a range of word forms—a property reflected in relative learnability. This increased reliability is shown in Table 2.13. These tables depict the patterns of violation of ALIGN incurred by a window stress system of this type. As the window size increases, a greater number of possible levels of violation need to be accounted for. This relative lack of reliability fails to distinguish such a large window system from many of its "neighboring" stress systems, yielding slower learning.

The question of maximal window size is less clear. These models do indeed predict some number of four-syllable windows, as shown in Figure 2.9. However, the predicted

	Alignment of designated property							
	0	1	2	3	4	5	6	7
Syllables	ALIGN violations							
2	0	**1**						
3	0	**1**	0					
4	0	**1**	0	0				
5	0	**1**	0	0	0			
6	0	**1**	0	0	0	0		
7	0	**1**	0	0	0	0	0	
8	0	**1**	0	0	0	0	0	0

(a) Patterns of violation of ALIGN with a same-edge default two-syllable window.

	Alignment of designated property							
	0	1	2	3	4	5	6	7
Syllables	ALIGN violations							
2	0	**1**						
3	0	**1**	**2**					
4	0	**1**	**2**	0				
5	0	**1**	**2**	0	0			
6	0	**1**	**2**	0	0	0		
7	0	**1**	**2**	0	0	0	0	
8	0	**1**	**2**	0	0	0	0	0

(b) Patterns of violation of ALIGN with a same-edge default three-syllable window.

	Alignment of designated property							
	0	1	2	3	4	5	6	7
Syllables	ALIGN violations							
2	0	**1**						
3	0	**1**	**2**					
4	0	**1**	**2**	**3**				
5	0	**1**	**2**	**3**	0			
6	0	**1**	**2**	**3**	0	0		
7	0	**1**	**2**	**3**	0	0	0	
8	0	**1**	**2**	**3**	0	0	0	0

(c) Patterns of violation of ALIGN with a same-edge default four-syllable window.

Table 2.13: Patterns of violations of ALIGN across window sizes. As the size increases, the amount of variability in violation also increases. The columns indicate potential positions for a designated property within a word. The rows give different word sizes.

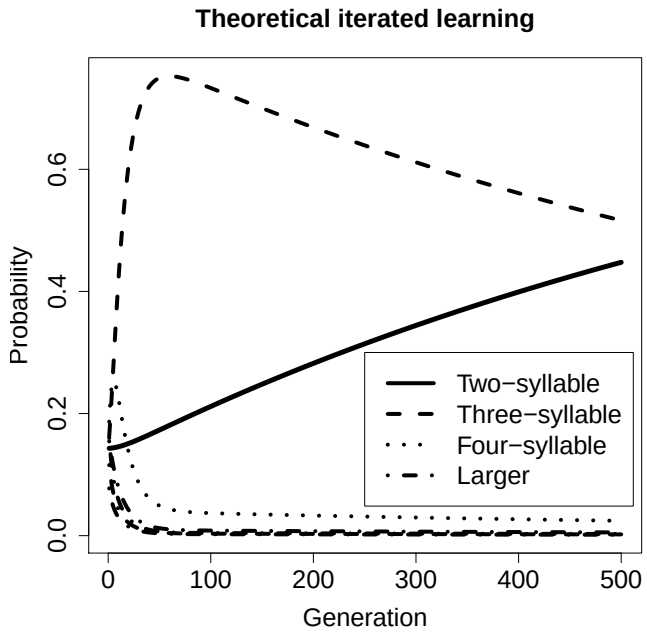


Figure 2.8: Dominance of small windows in predicted iterated learning. Proportion of stress windows taken up by a certain size across predicted generations. Calculated from learning results of Figure 2.6. $\eta = 0.1$.

probability of such a system is not necessarily high. This figure shows the predicted counts of various types of window stress, fitting the exponent of the predicted iterated learning to the data. Here the exponent was optimized by grid search from 1 to 10,000. With such a model, the expected number of four-syllable windows in a review such as the one Kager (2012) presents is 1.6. Due to the integer nature of count data, this prediction would be satisfied by only a single observation of a four-syllable window on either the left or right edge.

Is this prediction a success? This difference—a predicted one where a zero is observed—is difficult to evaluate statistically with an uninformative prior.[11] However, we can see that the observed data is highly consistent with the model. Although the fitted model predicts a value of one as most likely (32.69% probability), the

[11]We would essentially need to encode expectations about the probability of particular zeros. This is precisely the sort of tendency that is intended to be emergent in this approach.

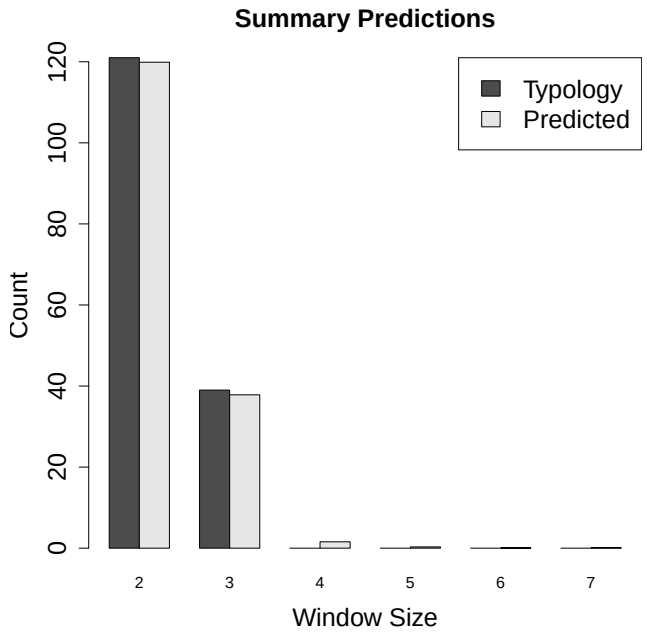


Figure 2.9: Comparison of predicted frequencies with typology (numbers from Kager, 2012). Counts sum over left and right windows. Exponent 1,664 chosen by minimizing sum squared error with data.

observed zero count is by no means unlikely (20.60% probability).[12] This is reassuring, but evaluation of the model based on the observation is less certain. This sort of issue is expected in a model in which learning predicts typology. In a traditional generative framework, typological gaps come in two types: accidental, and principled. The former are possible human languages which, for reasons of contingent historical accident, happen not to exist. The latter, in contrast, could not be human languages no matter what historical circumstances. Adding frequency subdivides the problem of gaps even further. These models can still predict zero frequencies if a pattern is simply not representable by a learner—categorical generalizations about typology are still possible. Thus principled gaps carry over in a simple way. However, frequency models

[12]Probabilities based on the resulting count for four-syllable windows in 10,000,000 random samples from a multinomial distribution with count equal to the size of the typology and probabilities equal to the model predictions.

will also predict small but non-zero probabilities for many patterns. If such a low probability language does not exist, this is not a simple "accident"—such situations are likely and expected under the model. The evaluation of a model rests on statistical measures of its accuracy, not just simple predictions about attestation or absence. Under this metric, the model succeeds at predicting a sufficient gap in probability between the likely two- and three-syllable patterns and the unlikely four-syllable one.

These simulations thus provide validation for a learning-based approach to explaining much of the tendencies in window stress typology. The frequency results are not obtainable with a traditional generative model, adding support for this kind of explanation. A low predicted probability for absent languages, in turn, addresses a criticism (Legendre et al., 2006) of Harmonic Grammar-like models—and quite likely other models which seemingly "overpredict" by allowing languages which are conceivable but difficult to learn.

2.4 Conclusion

In this chapter, I have presented several uses of a learning-based model of frequency in the typology of stress. Such models predict relative levels of attestation for stress languages on the basis of the distinctiveness of representations and the mutual reinforcement of stress data. These types of biases may be extracted from learning models in many ways; here I have focused on two: comparison of residual error and theoretical analyses of learning outcomes. The former gives a fairly direct measure of the relative speed of learning for different languages or patterns. The latter is useful in generation predicted typological counts, giving an estimate of the long-term probabilities of particular languages within a given set. This approach simplifies and exemplifies, but does not replace, models and experiments involving more direct measures of iterated learning.

I demonstrated the utility of such models of frequency using two problems in stress typology. The first problem, a correlation between parsing "direction" and primary stress location, highlights the ability of a learning-based model to make predictions even when the underlying grammatical framework does not. Languages in which primary stress does not depend on secondary stress or syllable count are more learnable due to their consistent representation—there is no need to exclude tendency-disobeying languages despite their infrequent occurrence. Such learning results require a suitable representation—here one with syllable-counting constraints referring to primary stress—thus forming an additional source of evidence for grammatical hypotheses. The second domain explored is window stress. There are distinct biases for small stress windows, including a typological gap above three syllables. I demonstrated that a learning-based model can account for these biases—again without forbidding infrequent patterns. The results on window stress show again that a learning account may address theoretical issues of overprediction—here problems arising from Harmonic Grammar's ability to count out arbitrarily large windows.

The kind of learning-based model presented here cannot hope to explain all frequency tendencies in typology. In this chapter I discussed only tendencies emerging from the mathematical structure of representations, not the phonetic substance. This is potentially the reason for some of the anomalies illuminated in both types of stress pattern. In §2.2.3 I suggest that the relative absence of patterns forming clashes at an edge with primary stress—despite their apparent formal learnability—might be due to perceptual problems with such patterns. Similarly, the inability of the window stress model to account for left/right asymmetries might owe credit to an external source of asymmetry between word edges. This kind of approach is explored, applied to nonfinality pressures, in the next chapter. Even these concerns do not exhaust the pressures on numerical typology. Languages do not simply grow "easier" over time. This is explained in part due to the probabilistic nature of learning: some-

times a failure to learn can make a language "harder," just by chance. However, this ignores more structured forces adding complexity to languages—borrowing, for example. These concerns are shared with categorical studies of typology. Despite the varied possible sources of bias on typology, this work lays out the potential power of considering learning itself as an immediate pressure on prediction, focusing on predictions arising from otherwise-motivated models of linguistic typology.

This work shows that taking learning seriously goes a long way toward accounting for frequency tendencies in typology. A grammatical theory—even a categorical one—plus a learning theory automatically yield a theory of typology in which learning biases can affect relative attestation. Here I focus on one particular grammatical framework (MaxEnt) and one particular learning algorithm (SGA), but the point is general. A learning theory is needed in any case, so considering frequency in this way adds no genuinely independent complexity to the system. Learning algorithms explicitly designed with this goal in mind (e.g. Riggle, 2008) will do well by a frequency-based criterion—but results presented here show that this is not obviously a necessary design feature. Ultimately, breaking down the wall separating grammatical formalisms from their learning yields a richer and more natural theory of language.

CHAPTER 3

INTERACTIONS BETWEEN LEARNING AND EXTRA-GRAMMATICAL BIASES

3.1 Introduction

The preceding chapters deal with biases that emerge from grammatical representations in the form of a Maximum Entropy constraint set. In this chapter, I consider biases that originate in properties external to that representation, focusing on a perceptual interpretation of nonfinality.

There are a number of ways in which stress typology is asymmetrical. Trochaic feet are more common than iambic ones. Penultimate and antepenultimate stress are well-attested, but peninitial and postpeninitial stress are much less so (e.g. Gordon, 2002). Left-to-right and right-to-left versions of otherwise identical patterns are not equally attested. These asymmetries are crucially *directional*: a sequence of stressed and unstressed syllables is differently attested when those syllables count from the left as opposed to when they count from the right. In this chapter I pursue a number of these biases with the goal of reducing them to a single fact about stress: the dispreference for stress on final syllables.

This type of directionality is different from the correlation discussed in Chapter 2 between the direction of secondary stress and the placement of primary stress. In that instance, the bias is toward primary stress systems placed in a particular place *with respect* to secondary stress. The directionality of secondary stress determines primary stress probabilistically in the typology, but that does not imply that secondary stress has its own *absolute* direction determined by learning biases. Indeed, in

the simulations presented in that chapter, no distinction is made between a pattern which proceeds left-to-right and one which proceeds right-to-left, so long as primary stress falls in the same location with respect to secondary stress in both instances.

In general, the biases of the previous chapters reflect a view of stress in which left is the same as right. This is fundamentally not the case, both in the actual observed typology of stress and in the demonstrated biases of human perception. This is to be expected: phonological strings are perceived, at least in part, in the single linear direction of increasing time. It would be surprising, at least, if the demands of this temporal extension on perception and production did not have any effect on stress languages. However, these numerical biases cannot emerge from learning alone in a constraint set that lacks directional biases. The abstract analyses of learning cares only for representations of stress forms, not their temporal properties. Thus any prediction for a leftward pattern must be true of the rightward reversed pattern—unless there exists at least one constraint which treats these types of strings differently *and* lacks a symmetrical counterpart.

One strategy to address numerical biases of directionality would therefore be to add symmetry-breaking elements to the constraint set, such as NONFINALITY with no corresponding NONINITIALITY. Individual constraints which are non-symmetrical are abundant, both in proposals on CON and in the preceding simulations. For example, alignment constraints obviously treat leftward and rightward strings in distinct manners—that is their principal purpose. However, if an alignment constraint to the right edge and an equivalent one for the left edge are *both* assumed, learning can use either of these symmetrically and the typological effects will be symmetrical. NONFINALITY differs in that it penalizes strings with final stress, but no constraint penalizing initial stress is necessarily assumed. This kind of asymmetry is by no means unique—for example, constraints on onsets routinely differ from constraints on codas.

I show that such a solution offers an effective model of a number of these biases, but I argue that this theory is incomplete. There are good reasons to believe that constraints like NONFINALITY are supported by pressures in (for example) perception and production (e.g. Gordon, 2000; Lunden, 2006). It would be desirable to explain typological biases in a way that connects with these extra-grammatical, semi-independent pressures. However, these factors do not in themselves tell the whole story: a particular perceptual bias might say that e.g. final stress is "weak," but that does not explain the resulting effect on what is actually found in the typology. To do this, a theory of learning is needed. A learning theory can explain why particular alternatives are likely in the face of difficulties in perception and production, and thus why particular "solutions" to these problems predominate.

In this chapter, I consider alternative interpretations of a nonfinality effect on the quality of input data. These include modeling this effect as additional "noise" on the final syllable.

3.2 Typological statistics of directional stress asymmetries

It is first necessary to establish the place of nonfinality as a pressure on stress typology. This is a part of the larger goal of understanding leftward/rightward asymmetries in the numerical typology. Though related, a claim for nonfinality as a typological influence is not exactly the same as a claim for its place among the constraints of CON, represented as NONFINALITY.[1] It is not as such crucial whether NONFINALITY is needed to represent particular languages in the typology; instead, the question is whether languages on the whole tend to stress strings in obedience of such a putative constraint.

[1]Throughout, *nonfinality* in typical case denotes the typological pressure (possibly phonetic) on typology. NONFINALITY in small caps denotes a particular constraint.

This question is not trivial. Languages are not typically wholly nonfinal or wholly final in their stress. Indeed, monosyllables offer an abundant counterexample: even many languages which typically do not allow final stress allow it in when final position is the only option available. More broadly, languages may override an apparent desire for nonfinality in favor of higher-ranked constraints, resulting in stress that is final or not depending on word length or the quantity of various syllables (see e.g. Prince and Smolensky, 1993/2004, Ch. 4).

Despite these difficulties, to examine the overall asymmetry of stress patterns at edges of the word, we must find *some* way to summarize the typology. Here I use Heinz's 2007 Stress Pattern Database as the sample for which to compute statistics. For 95 of the 109 patterns, Heinz's string generator was used to give all possible stress patterns from length two to length eight.[2] The strings generated included all possible weight distinctions of consequence to the languages.

Twelve n-gram properties of these strings were computed to summarize the directional asymmetries at edges, each of value 1 or 0. These are shown in Table 3.1. These features allow the discovery of asymmetries in basic patterns of one or two syllables on either edge of the word, abstracting away from the primary/secondary distinction and all weight distinctions.

For each language, the average value of these features was computed. In the calculation of an average, every assignment of a designated property to a given word size was taken to be equal to every other. Thus the average is an average over word shape *types*. Two perspectives were taken to the problem of word size. In the first, type frequency alone was used—equivalently, word size was uniformly distributed. In the second, word size was taken to decrease exponentially (base 2) with word length.

[2]13 languages were excluded because they did not generate with the tool used. These were languages with parentheses in their descriptions—no systematic bias should be expected from this exclusion. Pirahã was excluded due to the computational time required to generate all strings.

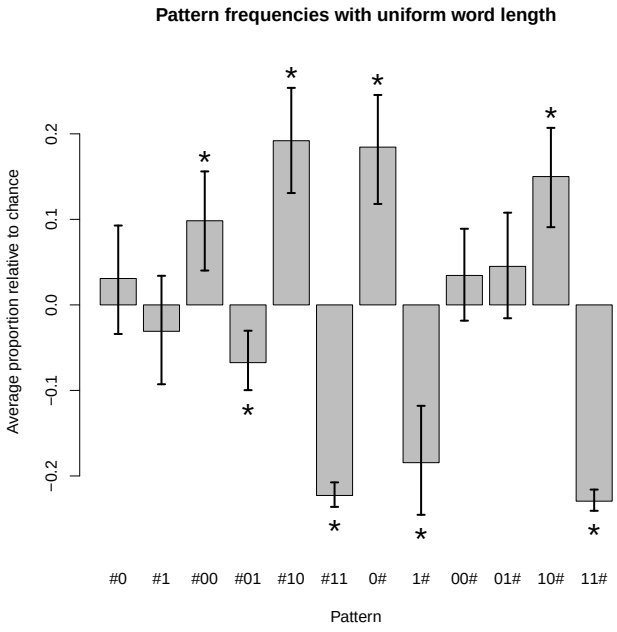
n-gram	Description
#0	Stressless first syllable
#1	Stressed first syllable (initial gridmark)
#00	Initial lapse
#01	Initial rise
#10	Initial fall
#11	Initial clash
0#	Stressless final syllable (nonfinality)
1#	Stressed final syllable
00#	Final lapse
01#	Final rise
10#	Final fall
11#	Final clash

Table 3.1: Edge n-grams used to summarize typology.

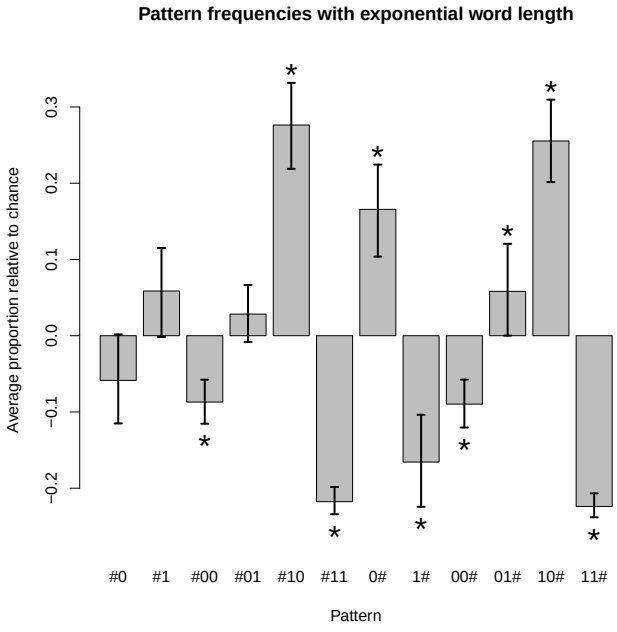
Each set of features was weighted by the frequency of that stress pattern in Heinz's Stress Pattern Database. Put another way, the feature list for each language is the probability-weighted sum of the lists for each word length, which are in turn the raw means of each word shape of the corresponding length.

In Figures 3.1a and 3.1b, these averages are compared with assumed "chance" values. Chance here is as predicted by uniformly random stress assignment to syllables: 0.5 for single-syllable patterns on word edges, 0.25 for two-syllable patterns. This assumption is somewhat implausible: first, stress probability is influenced by adjacent context, as we well know; second, if a language is required to have at least one stress per word, this further influences chance assumptions. However, this view of the chance rate of stress assignment suffices on a coarse level.

Comparisons are marked as significant if there is a probability less than criterion that the difference between the observed frequency and chance includes 0.0, calculated by 10,000 bootstrap resamplings of the data. Comparisons are made for each of the 12 features for two distribution assumptions, with six additional pairwise comparisons. Significance is therefore indicated at $\alpha = 0.05$, so the Bonferroni-corrected two-tailed criterion for these 30 tests is $(\frac{1}{2})(\frac{0.05}{30}) = 8.3 \times 10^{-4}$.



(a) Assuming only the type frequency for distinct word lengths.



(b) Frequency of word lengths assumed to be exponentially decreasing.

Figure 3.1: Frequency of n-gram patterns relative to chance in the languages of Heinz (2007). Error bars indicate 95% interquartile range of bootstrap. See text for discussion of chance and significance.

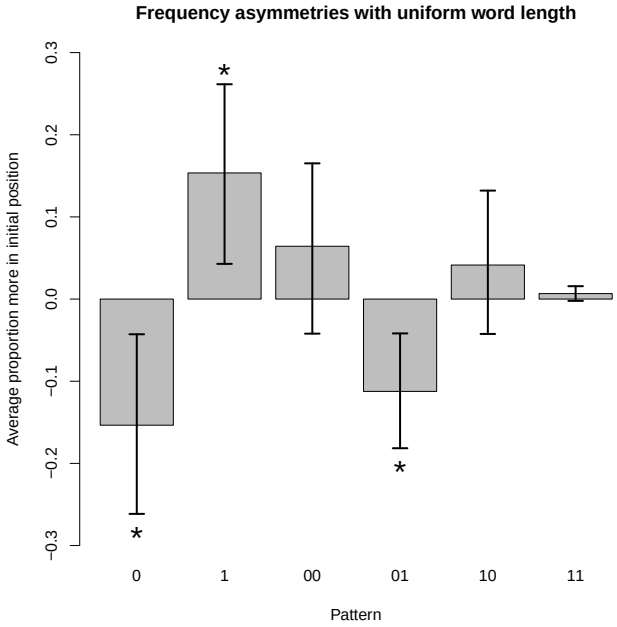
We can focus attention further on features which are significant under both distributional assumptions and which deviate from chance in the same direction for both sets of tests. These n-gram features are shown in Table 3.2. This table shows a particular n-gram that is found to be significant in the statistical test and gives its deviation from chance. For example, initial falls are trigrams involving the left edge, a stress, and an unstressed syllable. This trigram deviates significantly from chance, being higher than expected. It is therefore *favored*. In contrast, initial clash is *disfavored*.

n-gram	Description	Tendency relative to chance
#10	Initial fall	favored
#11	Initial clash	disfavored
0#	Stressless final syllable	favored
1#	Stressed final syllable	disfavored
10#	Final fall	favored
11#	Final clash	disfavored

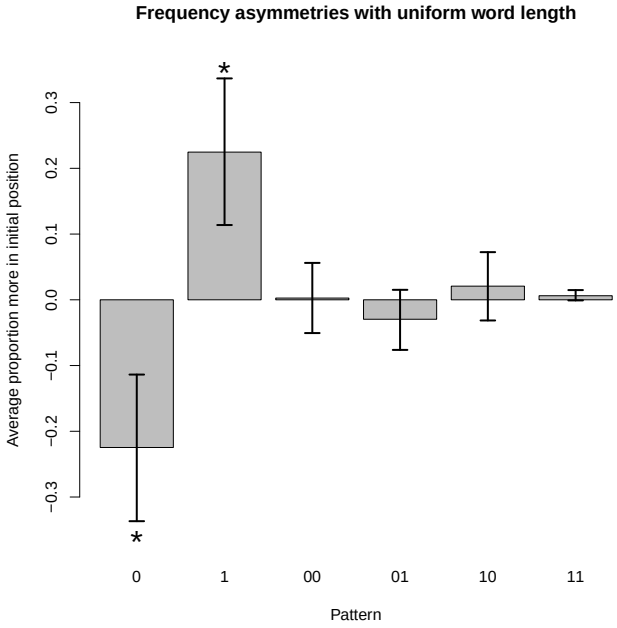
Table 3.2: Edge n-grams significant under both assumptions with equal directions of deviations from chance.

Some of these comparisons point toward typical phonological observations. For example, clash is disfavored on either edge but stressless edge syllables seem to be preferred on the right edge. To take this further, it is necessary to compare the average proportions of words with these features on the left edge with their counterparts on the right. This comparison is shown in Figures 3.2a and 3.2b. Significance was determined as before.

A single stressed or unstressed syllable on the word edge is relatively favored significantly under both distributional assumptions. On the left edge, stress is favored. On the right, stressless syllables are favored. These two trends both have theoretical representations as, for example, NONFINALITY(Syllable) and INITIALGRIDMARK (Hyde, 2008), respectively.



(a) Assuming only the type frequency for distinct word lengths.



(b) Frequency of word lengths assumed to be exponentially decreasing.

Figure 3.2: Frequency of n-gram patterns at the left edge of a word compared to the right. Positive difference indicate a bias for the left edge, negative differences for the right. Bars indicate 95% interquartile range of bootstrap. See text for discussion of chance and significance.

These two features are correlated. In a linear regression predicting initial stress from final stresslessness(within all strings analyzed for the Stress Pattern Database languages, the coefficient is significant for both the uniform distribution (Estimate: 0.48, $p < 0.05$, $r^2 = .22$) and exponential (Estimate: 0.69, $p < 0.05$, $r^2 = .54$).

These results suggest that asymmetries between the right and left edge are real. In principle, the underlying factor driving this difference could be left edge bias, right edge bias, or both. Due to the statistical tradeoff between the two edges, it is more parsimonious to—initially, at least—pursue an account in which only one of the trends is actively motivated, explaining the other as its secondary consequence. In the following section I discuss some reasons to believe that nonfinality is perceptually motivated and thus should be added first to the account of numerical typology.

3.3 Motivations for nonfinality

Several accounts have been given for the origins of nonfinality as a typological pressure or phonological constraint. I discuss three of these in turn.

3.3.1 Tonal crowding

Gordon (2000) (following Hyman, 1985) proposes that NONFINALITY the constraint and nonfinality the pressure on languages both arise from an effort to avoid tonal crowding. He observes that many languages possess final boundary tones. Initial boundaries are, in contrast, less common and (perhaps) less dramatic. In a language that has a final tone marking a phrase-level prosodic boundary, he reasons, a word-final stress should be undesirable. In such a case, the final phrasal boundary tone and the tonal associate of stress will be pressed to the same syllable, resulting in a contour tone. There is an apparent cross-linguistic preference to avoid contour tones (i.e. distinct tonal targets on a single syllable), and thus there should be a preference to avoid an overlap between final stress and phrase boundary tones.

This account drives the typology of stress with something outside the immediate grammatical mechanisms responsible for stress assignment, placing responsibility on the typology of tonal boundary marking. While it does indeed appear that final boundary tones are more common, this in itself is not explained. Thus the explanatory onus is simply pushed higher in the prosodic hierarchy. However, this is not in itself a problem for the account. More problematic (as Gordon notes) would be if there were no correlation between boundary type/location and nonfinality. However, even this need not be an issue if the general perceptual pressure yields a universal constraint NONFINALITY and no NONINITIALITY.

Another issue with this account is that it assumes stress is principally tonal in nature. There are many correlates of stress, with seemingly none of them shared by all languages. This reliance is therefore troubling, though not damning.

3.3.2 Clash avoidance

A second explanation for nonfinality is the avoidance of clash (Gordon, 2000; Karvonen, 2008). This approach is more internal to stress assignment, relying only on the distribution of other stresses. The fundamental observations are that clash is dispreferred cross-linguistically and that it is not necessarily word-bounded. Final stresses will create clashes with following initial stresses, while penultimate stress can never do this (Tables 3.3 and 3.4). There is a general pressure for stresses to be close to edges, either for demarcation or—as discussed in previous chapters—for learnability reasons. This account thus frames the typology as "edgemost if no clash, else next edgemost."

Clash	Clash avoided by nonfinality
$\sigma\grave{\boldsymbol{\sigma}}\#\grave{\boldsymbol{\sigma}}\sigma$	$\grave{\sigma}\sigma\#\grave{\sigma}\sigma$
$\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma\grave{\boldsymbol{\sigma}}\#\grave{\boldsymbol{\sigma}}\sigma\sigma$	$\grave{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma\#\grave{\sigma}\sigma\sigma$

Table 3.3: Nonfinality removes clash with following initial stress.

Clash from finality	Clash from nonfinality
$\sigma\grave{\sigma}\sigma\boldsymbol{\grave{\sigma}\#\grave{\sigma}}$	$\grave{\sigma}\sigma\grave{\sigma}\sigma\#\grave{\sigma}$
$\grave{\sigma}\#\sigma\grave{\sigma}\sigma\grave{\sigma}$	$\boldsymbol{\grave{\sigma}\#\grave{\sigma}}\sigma\grave{\sigma}\sigma$

Table 3.4: Nonfinality can create clash with preceding final stress.

This account is attractive in particular due to monosyllables. These words will—if stressed at all—always have initial stress. Therefore any word obeying nonfinality and preceding a monosyllable will avoid clash. Short words are overwhelmingly abundant in stress data, and so this seems like a potential large markedness savings. Where the account falters in this respect, however, is when we consider preceding words. A final-stress form such as a monosyllable *preceding* a form obeying nonfinality can have clash across boundaries. In the typically common disyllables, penultimate stress is equivalent to initial stress. It will therefore clash with any preceding final stress.

This conflict can only be arbitrated by the relative frequency of final and initial stresses. If initial stress is more common than final, the clash account will come out in favor of nonfinality. This is exactly what we observed in the preceding statistical analysis. However, this is then an account of nonfinality in terms of initiality—the question regresses to why initiality should be common.

3.3.3 Final lengthening

A third type of explanation for nonfinality is based in a phonetic effect unique to final position: final lengthening (Lunden, 2006). This approach echoes the tonal crowding account, using length rather than tone. Final position is subject to near-universal phonetic lengthening for at least some levels of prosody (e.g. Oller, 2005). The final syllable will be longer than a similar syllable elsewhere in the word, all else being equal. This being the case, the final syllable might not be a good host for stress cue; the lengthening from stress would be difficult to separate from the lengthening due to position.

This account has the advantage that it attaches the asymmetry to an established asymmetry in production, namely final lengthening. Final lengthening itself requires explanation, but this is not an added burden. Similar to Gordon's approach, the account has little to say about stress systems in which length is not an important cue.

3.3.4 Summary

These three accounts have very different ways of approaching nonfinality. However, each is an attempt to connect a nonfinality pressure to something outside the immediate auspices of stress and its assignment. All these approaches share something else: they determine something undesirable about final stress, but do not necessarily determine the alternative to such systems. That is, though they state what should be *avoided*, there is no immediate answer to the question of what should be done instead. Superficially, this is similar to concepts like markedness constraints in OT. However, the impact and consequences of these constraints are understood in the wider framework of an OT grammar; what is lacking in these phonetic explanations is such a framework.

In the next section I elaborate models incorporating learning and types of nonfinality biases based on NONFINALITY and accounts such as those based in tonal crowding and final lengthening. These simulations offer insight into the connection between biases on individual syllable stressedness and the frequency of stress patterns reflecting these biases. I show that such methods allow concrete exploration of the options open to a learner subjected to a disfavored final stress.

3.4 Nonfinality as external bias on learning

3.4.1 Simulating nonfinality

In this chapter I discuss two distinct types of approach to modeling a nonfinality bias on learning and typology: biases based on assumptions about constraints and biases based on assumptions about the quality of input data.

The simplest type of assumption about constraints concerns whether a given type of constraint is contained within Con or not. This is the typical *modus operandi* of categorical typological predictions, and can extend reasonably to predicting frequencies. For nonfinality, these types of assumptions concern what constraints are available that penalizes stress near word edges. These could be Nonfinality referring to syllables or to feet, or Noninitiality echoing these constraints at the left edge. Further afield, such constraints could be assumed to differ in their relative starting weights or learning rates.

These approaches at best indirectly encode intuitions drawn from the possible origins of a nonfinality effect. Indeed, this is a concern with OT constraints in general—even if their putative phonetic motivations are clear, the relationship between such grounding and their formal status is unclear (but see Hayes, 1999; Smith, 2004, on formalizing phonetic grounding). These concerns motivate the examination of the other class of biases, based instead in attempts to model the origin of the effect itself.

Here I simplify the theories of nonfinality effects, abstracting them to a simple notion that there is some bias for stress to be missed—or ignored—by the learner when it would create final stress forms. This bias has a natural interpretation within the learning framework discussed in this work: input data, normally sampled from the teacher and provided to the learner, is not always given to that learner unaltered. Instead, the "channel" affects the learner's simulated perception of the form, giving it non-veridical forms as targets for learning.

In the most straightforward interpretation, strings passed to the learner with final stress have that stress removed. Such a change can affect the ease of learning while still permitting, for example, a fixed final stress language in *some* constraint sets. For example, the n-gram set discussed in Chapter 1 can encode the occasional misperceived string as "erroneous" weight on a final n-gram lacking stress. This is not true, however, for constraint sets such as the foot-based one pursued in Chapter 2. Most constraints on feet and stress cannot be violated in the absence of any foot structure. For a destressed final stress datum to be informative, then, the misperception must not only remove stress, but also place it in a new location.

The correct position for misperceived stress is not immediately obvious. As the most neutral approach, stress could be randomly reassigned to any unstressed syllable of the form in misperception. In an approach grounded more closely in, for example, the tonal crowding account of nonfinality, final stress could be misperceived as stress on the penult instead. Finally, in the presence of final stress, the learner could choose to disregard the input and instead learn on the output of its own grammar, consolidating on its already-accrued information about its target language. This choice turns out to not be crucial in most cases. Most words are short, so reassignment of stress—if non-vacuous—will very often shift stress from the final syllable to the penult. This is because in short words there are few (or no) other options. In the following, I consider each of these options. These three types of bias simulation are explicated in Figure 3.3.

There is some support for these kinds of pressures in synchronic phonology. For example, Axininca (McCarthy and Prince, 1993b, pp. 159–160) exhibits a pattern similar to the "penult" method of stress reassignment. Axininca stress is left-to-right iambic. However, stress cannot be final. Thus in (for example) even-parity forms consisting of light syllables, there is a conflict between iambicity and nonfinality. This is resolved in one of two ways: deletion or penult stress. The latter therefore involves

Given datum x has final stress, with probability p do the following:

Penult	Assign stress to the penult, then learn on that new datum (disregarding the original one).
Randomize	Assign stress to a randomly selected syllable, then learn on that new datum (disregarding the original one).
Reparse	Discard the original datum. Parse the word shape according to the *learner's* current grammar. Learn on this output.

Figure 3.3: Description of three possible probabilistic approaches to modeling a nonfinality perceptual bias.

iráawanàti	'su caoba'	left-to-right iambic, nonfinal
kimítaka ∼ kimítàka	'quizá'	reversal or deletion
máto	'polilla'	mandatory reversal

Table 3.5: Axininca avoidance of final stress (McCarthy and Prince, 1993b, pp. 159–160).

a local reversal of the apparent foot shape from iambic to trochaic. In disyllables this reversal is mandatory. This option in Axininca stress is given a historical grounding if the stress reassignment approach is valid.

3.4.2 Fixed stress

Directional asymmetries are most obvious—and most circumscribed—in the typology of fixed stress. Here, as before, I focus on fixed stress systems with a single stress. It is here, therefore, that I begin testing this approach to integrating nonfinality pressures with typological prediction. I first discuss generalizations from the Stress Pattern Database, shown in Table 1.3.

In fixed stress systems, final and initial stress are roughly at parity in their frequency. There is no convincing support for the abundance of one over the other. From this typology, at least, any explanation from non*finality* must actually accommodate

	From left	From right
	σσσσσσσσ́	*σσσσσσσ́*
Distance 0	initial	final
	69 languages	74 languages
	σσ́σσσσσ	*σσσσσσ́σ*
Distance 1	peninitial	penultimate
	12 languages	60 languages
	σσσ́σσσσ	*σσσσσ́σσ*
Distance 2	postpeninitial	antepenultimate
	0 languages	8 languages

Table 3.6: Fixed stress languages with counts from Heinz's (2007) Stress Pattern Database. Reproduction of Table 1.3

finality. Instead, we see that right-edge patterns away from the edge—penultimate and antepenultimate—are overattested in comparison to their mirror-image patterns at the left edge—peninitial and postpeninitial.

The typology of fixed stress is complicated in that not all surveys agree. In Table 3.7, four surveys are shown encompassing edge stress: Hyman (1977), Gordon (2002), Heinz (2007)[3], and Goedemans and van der Hulst (2013). These surveys are not uniform in their methodology. For example, Hyman includes non-fixed stress while the other counts do not. Heinz incorporates Gordon's survey as well as one of Bailey (1995). The Goedemans and van der Hulst numbers are those presented in the *World Atlas of Language Structures* chapter extending StressTyp.

Despite the variety of approaches, some facts are clear from these surveys. They all support the relative prevalence of initial stress, the rarity of peninitial and antepenultimate stress, and the marginality of postpeninitial stress. Where the surveys differ is with respect to final and penultimate stress. All studies shown, except WALS, support the generalization given above: final and initial stress are roughly equally common, with penultimate stress next common. WALS instead reverses the order

[3]These numbers are derived from Heinz's Stress Pattern Database and not directly presented in the work cited.

	Hyman		Gordon		Heinz		WALS	
	#	%	#	%	#	%	#	%
initial	114	37.3	61	30.8	69	30.9	92	32.6
peninitial	12	3.9	12	6.1	12	5.4	16	5.7
postpeninitial	0	0.0	0	0.0	0	0.0	1	0.4
final	97	31.7	63	31.8	74	33.2	51	18.1
penultimate	77	25.2	55	27.8	60	26.9	110	39.0
antepenultimate	6	2.0	7	3.5	8	3.6	12	4.3

Table 3.7: Comparison of typologies of edge stress in Hyman (1977), Gordon (2002), Heinz (2007), and Goedemans and van der Hulst (2013).

of final and penultimate stress, placing the frequency of penultimate as greater than twice that of final stress and predominant among stress patterns. This difference is concerning, but I will not arbitrate it directly. Instead, I will show that empirical distinctions such as this can be accounted for by changes in assumed parameters (*viz.* misperception probabilities) modeling nonfinality as a perceptual effect. These parameters should ideally be derived empirically in any case, so the fact that our typological knowledge cannot decide the question yet is not overly worrying.

In the WALS data, the nonfinality bias is more readily apparent than in the other typologies. WALS finds an outright reversal in favor of penultimate stress compared to what would be predicted based on distance from the word edge alone. However, it is not the case that bias is absent in, for example, the Stress Pattern Database. As mentioned in Chapter 1, distance from the word edge and edge of alignment are not independent in the Stress Pattern Database ($\chi^2 = 25.39$, $p < 0.05$). This test allows us to see how individual patterns differ from what would be expected under independence. The expected counts are shown in Table 3.8. These counts suggest that final stress is observed less than expected, while penultimate stress is observed more than expected—a nonfinality bias.

The simulation methods discussed in §3.4.1 all serve to do one thing: they render patterns involving final stress more difficult to learn accurately by simulating mis-

	From left	From right
'Distance 0	52 languages	91 languages
Distance 1	26 languages	46 languages
Distance 2	3 languages	5 languages

Table 3.8: Rounded expected counts from a χ^2 test of Table 3.6: $\chi^2 = 25.39$ $df = 2$, $p < 0.05$. Replication of results from Table 1.4.

perception. This is because these methods "scatter" data on final-stress forms across the space of violation vectors. If the learner is to acquire its teacher's language exactly, it must necessarily disregard some of the data it receives which misplaces final stress. These misperceived strings slow the learner's progress toward confidence in a hypothesis or shift it into a new interpretation of its learning data.

Of the six canonical fixed stress systems, only three are affected by reassigning final stress. These are final, peninitial, and postpeninitial stress. The point is obvious with final stress: misperception distorts the one and only stress of the word every time it occurs. Peninitial and postpeninitial stress acquire their disadvantage in short strings. A stress that counts away from the left edge will be found in final position of the word is short enough. Thus peninitial stress appears final in words up to length up to length two, postpeninitial for words up to length three. This seems negligible, but is actually dramatic when we consider the skew in the distribution of word lengths toward quite short words. These two systems are thus quite affected by any alteration to final stress.

The typology of fixed stress systems is the product of two different pressures. Learning bias emerging from distinctiveness and reliability, as discussed in previous chapters, creates a pressure for stress closer to word edges. This accounts for the high frequency of final and initial stress and the relative scarcity of, for example, antepenultimate stress. A nonfinality pressure for misperceiving final stress reduces the frequency of peninitial and postpeninitial systems, mostly to the benefit of penul-

timate stress. These pressures, properly construed, could account for the numerical patterns of Table 3.6.

3.4.2.1 Choosing a constraint set

It is not possible to model the set of canonical fixed stress languages with the constraint set discussed in Chapter 2. The existence of antepenultimate stress necessitates some constraint which will be violated *less* by candidates which are inexactly aligned with the right edge of the word. A clear choice for such a constraint is a familiar one, NONFINALITY (Prince and Smolensky, 1993/2004). This is a constraint which assigns a violation to candidates with "final" stress, for some definition of final. However, NONFINALITY has a number of alternate interpretations. Of principal importance is the distinction between NONFINALITY(Syllable) and NONFINALITY(Foot). The former penalizes only literally final stress, while the latter also penalizes final feet.

If nonfinality effects are to be modeled as emergent from factors other than assumptions about CON, it would be ideal to assume constraints which are symmetrical at left and right edges. Thus we are motivated to consider NONINITIALITY constraints. This is in addition to arguments for noninitiality in its own right, largely divorced from fixed stress typology (e.g. Alderete, 1995; Kennedy, 1994; Kenstowicz, 1993). If a NONINITIALITY constraint is to be considered, it presents an issue echoing the one with NONFINALITY: should there be NONINITIALITY(Syllable), NONINITIALITY(Foot), or both?

One approach is to choose a constraint set based on performance. In Chapter 2, I discussed how to fit the exponent in simulated iterated learning to the observed typological data, minimizing sum squared error. Probabilities are computed by tracking the resulting learner languages for a range of teacher languages. This data is exponentiated to best fit the typology. This procedure picks the best attained prediction of

frequencies for a particular configuration of learning assumptions. We can repeat this procedure for the fixed stress typology, varying only assumptions about the constraint set.

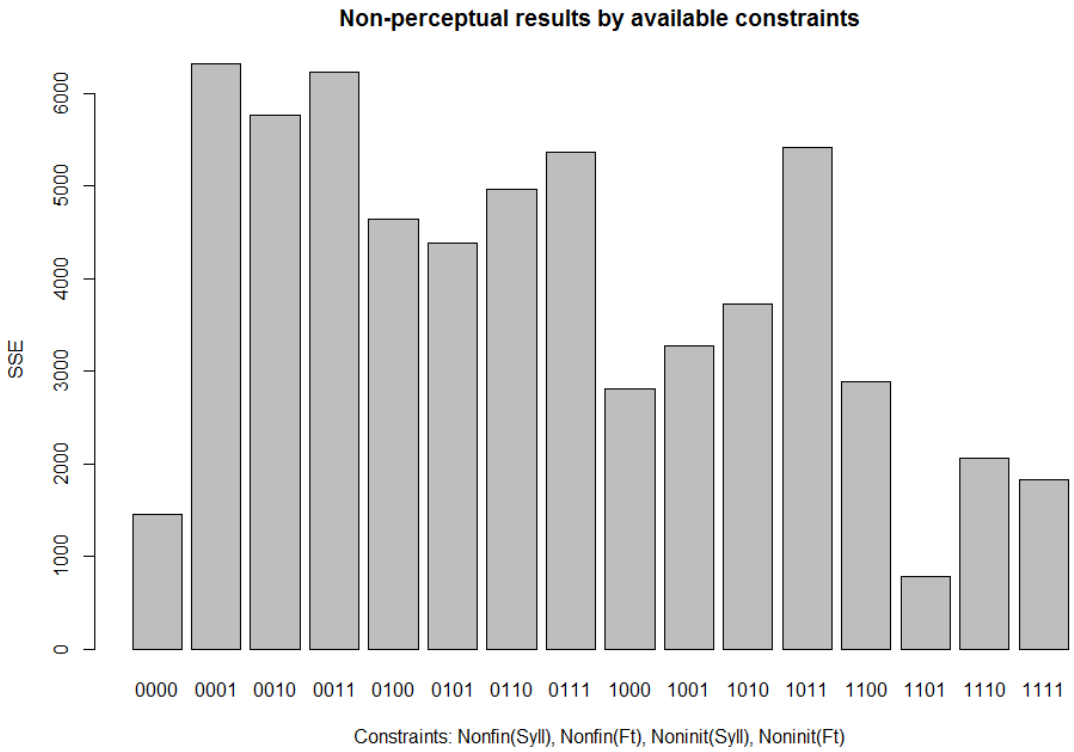


Figure 3.4: Best performance at typological frequency prediction for fixed stress across assumptions for NONFINALITY and NONINITIALITY. Codes read with 1 for presence of a constraint, 0 for absence in the following order: NONFIN(Syll), NONFIN(Ft), NONINIT(Syll), NONINIT(Ft).

Results are shown in Figure 3.4. A fairly wide range of errors is found within the possible range. The minimum possible error is 0, with exact prediction of the frequency of every language in the set. This is the ideal, at the bottom of the plot. There are 223 languages in the sample, so the maximum error is $223^2 = 49,729$. Thus the scale of the graph occupies about 12% of the total range. On the right half of the results we see results for constraint sets with NONFINALITY(Syll), which are largely good. These are indicated by an initial 1 in the axis labels. Two standouts are clear.

The first is a constraint set with *no* NONFINALITY or NONINITIALITY constraints, illustrating the relative size of the (here, unexplained) directional asymmetry compared with the primacy of word edges found in learning generally. The second is a constraint set with *both* NONFINALITY constraints and only NONINITIALITY(Ft). Hearteningly, this NONINITIALITY is the one argued for when this constraint is discussed (e.g. Kennedy, 1994).

These results point to two views of modeled nonfinality effects. First, we can assume the fully unbiased constraint set, with all four constraints. With this approach, all asymmetries are attributable to the modeled nonfinality. Second, we can use the best constraint set found here, which lacks NONINITIALITY(Syllable). I discuss both in the following.

3.4.2.2 Simulation results

Figures 3.5 and 3.6 show the best obtainable typological prediction error for given values of the probability parameter and given methods of simulating a nonfinality pressure. For each combination of a method and a probability, simulations are performed to calculate a transition matrix as discussed in Chapter 2. This transition matrix records the probability that a learner will acquire any given language within a set when its teacher possesses some (possibly different) language in that set. Here the set of languages is simply the fixed stress languages.

As previously discussed, these transition matrices can lead to longer-term typological predictions by exponentiation. Here, I attempt to show how *compatible* a given set of learning results is with the observed typology. To do this, the exponent of the matrix is optimized[4] with the objective that error between the observed typology and the predicted probabilities is minimized. In doing this, we give the learning approach its *best possible* chance of success. If there exists any exponent which makes the re-

[4]Optimization is with a simple grid search up to generation 10,000.

sults agree with typology, it will be discovered. This means that the raw prediction results are not of primary interest. Instead, we seek to compare the performance of both different nonfinality simulation methods and different values of the probability parameter.

Between these two, the probability parameter is the more crucial. If the method succeeds at explaining the typology better than *not* simulating nonfinality, we expect to see a decrease in error at some point greater than probability zero. This reflects an improvement when the method is "turned on." In addition, we might have *a priori* suspicion that very high parameter values would be disadvantageous—these states reflect a situation in which learning is entirely subject to the whims of the nonfinality pressure. As we know these pressures are not categorical (at least as the methods are construed here), this would not seem appropriate. Taking these two in combination, we might then expect that an effective nonfinality simulation method should show a local minimum in error at some point greater than zero and less than one.

Figures 3.5 and 3.6 show just such an effect for fixed stress using the WALS typological counts. The penult and random methods both exhibit local minima somewhere in the 0.10–0.20 range. It is unsurprising that these two methods should perform similarly—they both involve reassigning stress on the receipt of a final-stressed form. In many cases, the syllable stress lands on will be the same in either method. Randomization of stress in a two-syllable word will either be vacuous (i.e. it will be placed back on the final syllable) or it will be penultimate. In three-syllable forms, this remains likely even if there is another option (initial stress). Because there is a bias toward small words, we can see that most of the time these two biases will be doing very similar things. In contrast, reparse does not seem to have a minimum, rather leaving no real effect on the best typological predictions. This shows, minimally, that not all conceptions of a bias will exert similar forces on typology.

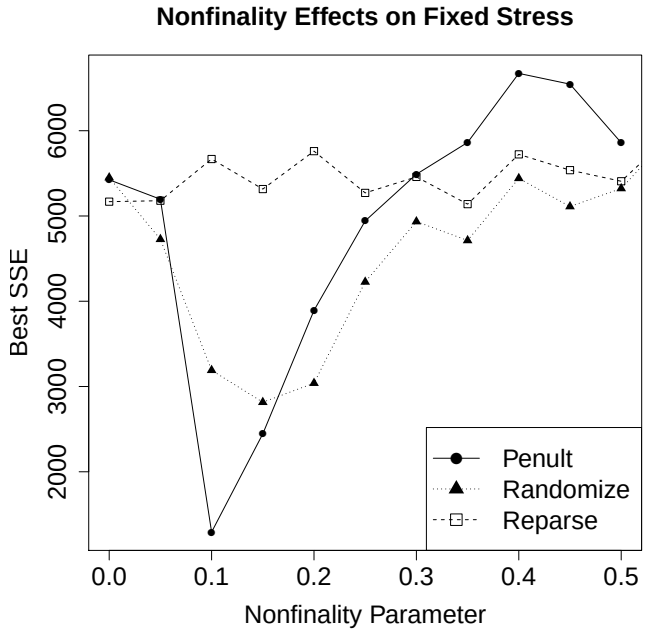


Figure 3.5: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. NONINITSyll included, typology from WALS. $\eta = 0.1$.

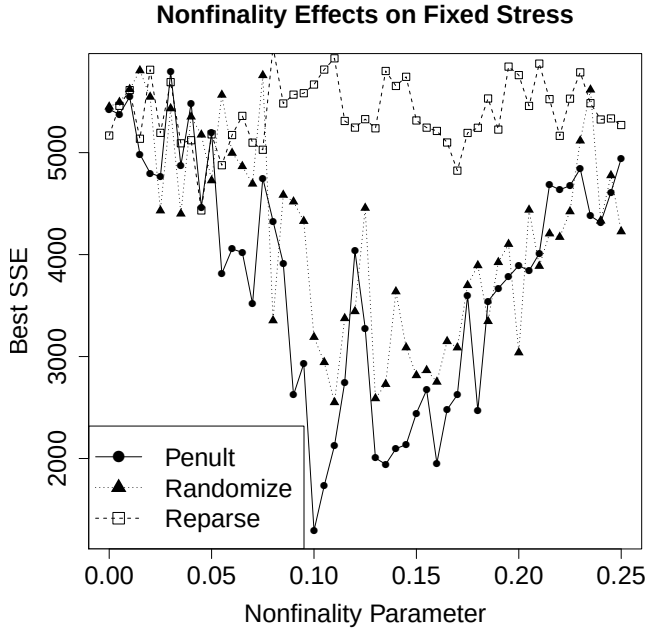


Figure 3.6: Probability optimization over a restricted range. NONINITSyll included, typology from WALS. $\eta = 0.1$.

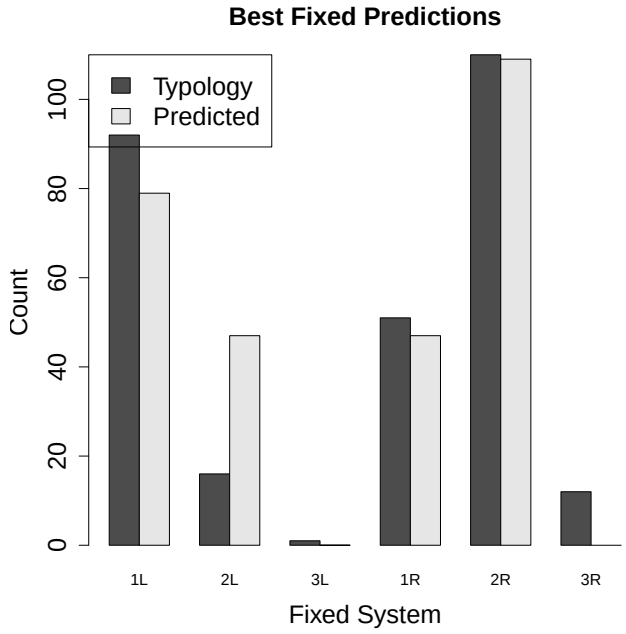


Figure 3.7: Typological predictions with best results from optimizations: Penult stress reassignment with probability 0.10 and 1290 generations. NONINITSyll included, typology from WALS. 1L, 2L, 3L mean fixed stress on the first, second, and third syllable from the left; symmetrical for 1R, 2R, 3R on the right.

In Figure 3.7, I show the best results according to these methods. We see here that the model is able to capture the asymmetry between final and initial stress and the relative advantage of penultimate stress in the WALS dataset. However, some types of language are not accurately modeled, even if the overall error is decreased. In particular, peninitial stress is overpredicted and antepenultimate stress is underpredicted. This is consistent with the model valuing general learnability over other potential concerns, and thus privileging the bias to have small distances from the word edge.

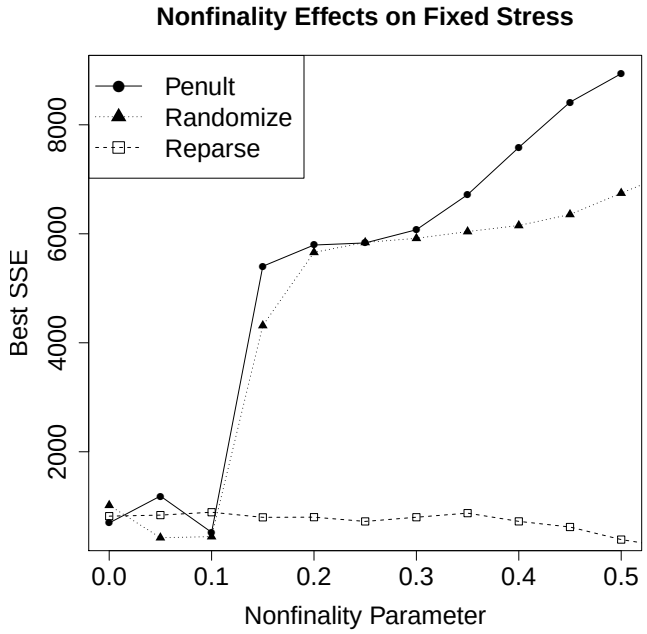


Figure 3.8: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. NoninitSyll not included, typology from WALS. $\eta = 0.1$.

Figures 3.8 and 3.9 show results when the constraint set is selected to exclude Noninitiality(Syllable) as discussed in §3.4.2.1. In this case, a local minimum is far less obvious. There is an apparent minimum above zero, but it is hard to verify and definitely less reliable than the one shown previously. However, the actual count predictions in Figure 3.10 show results that are not much-degraded beyond the ones

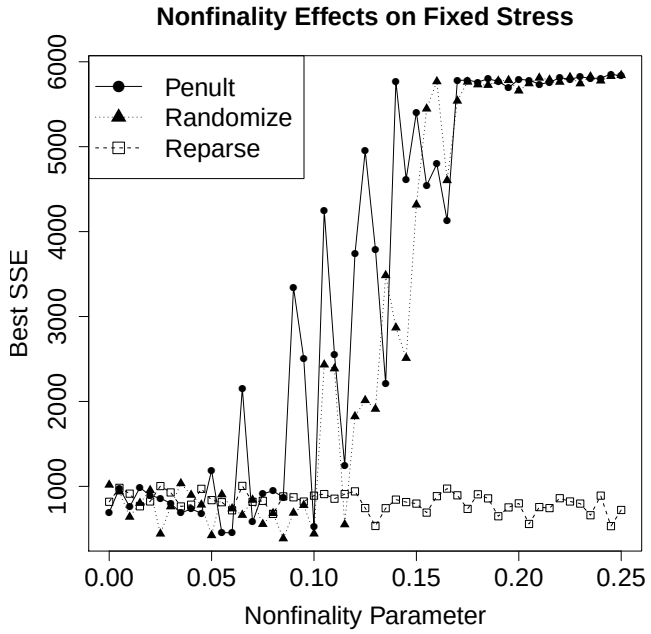


Figure 3.9: Probability optimization over a restricted range. NONINITSyll not included, typology from WALS. $\eta = 0.1$.

we saw before. What is happening here? The choice of constraints essentially trades with the simulated nonfinality effect—constraining the constraint set to work without NONINITIALITY(Syllable) works similar effects to simulating a perceptual issue with the right edge. This result reinforces the idea that frequency results are *contingent on grammatical assumptions*, even when we introduce supposedly extra-grammatical biases. Here, that additional bias can do little more than is already done by the structure of the constraint set, rendering the model less effectual in this regard.

We may now turn our attention to modeling the other view of fixed-stress typology, represented by the counts from the Stress Pattern Database. Recall that the primary distinction here is in the place of penultimate stress—is it privileged, as in WALS, or behind final stress, as in the Stress Pattern Database? Simulations with an unbiased constraint set are shown in Figures 3.11 and 3.12. These results again show a less dramatic minimum for the nonfinality simulation. The relative inadequacy of this model is further shown in Figure 3.13. The Stress Pattern Database has a lower

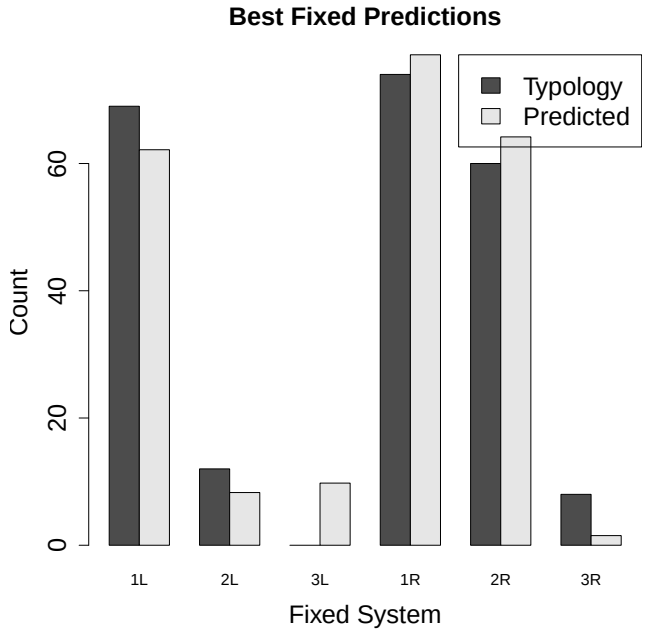


Figure 3.10: Typological predictions with best results from optimizations: Reparse stress reassignment with probability 0.60 and 169 generations. NoninitSyll not included, typology from WALS. 1L, 2L, 3L mean fixed stress on the first, second, and third syllable from the left; symmetrical for 1R, 2R, 3R on the right.

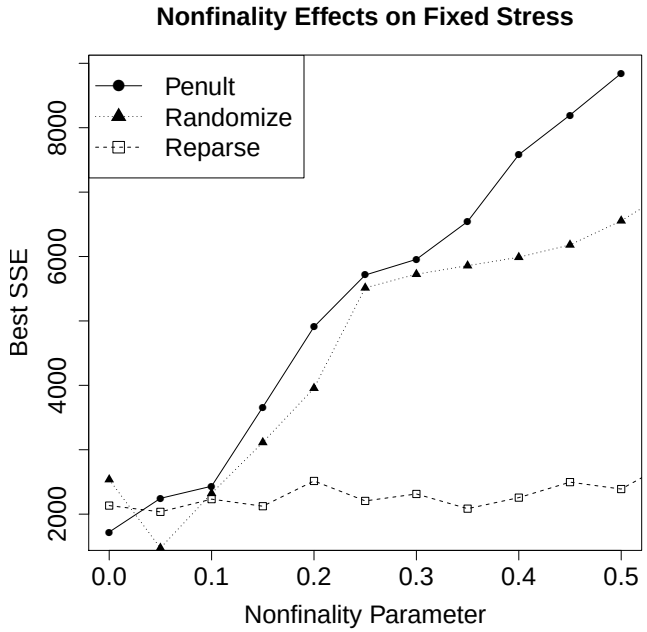


Figure 3.11: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. NONINITSyll included, typology from the Stress Pattern Database. $\eta = 0.1$.

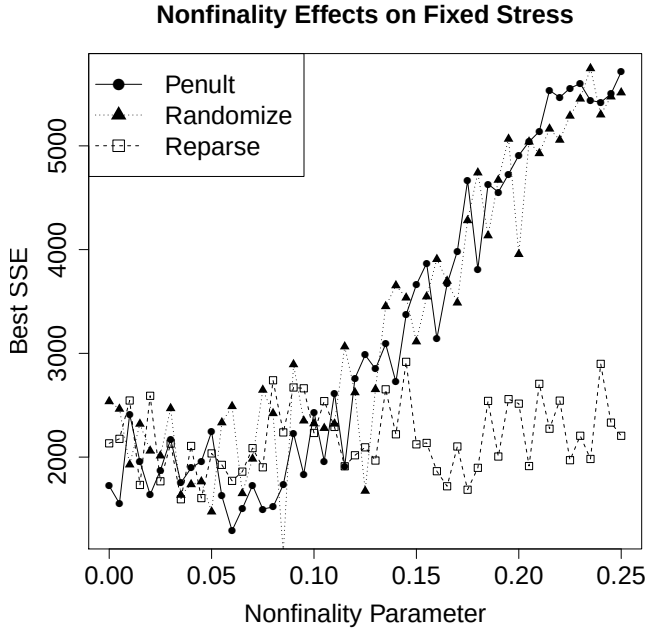


Figure 3.12: Probability optimization over a restricted range. NONINITSyll included, typology from the Stress Pattern Database. $\eta = 0.1$.

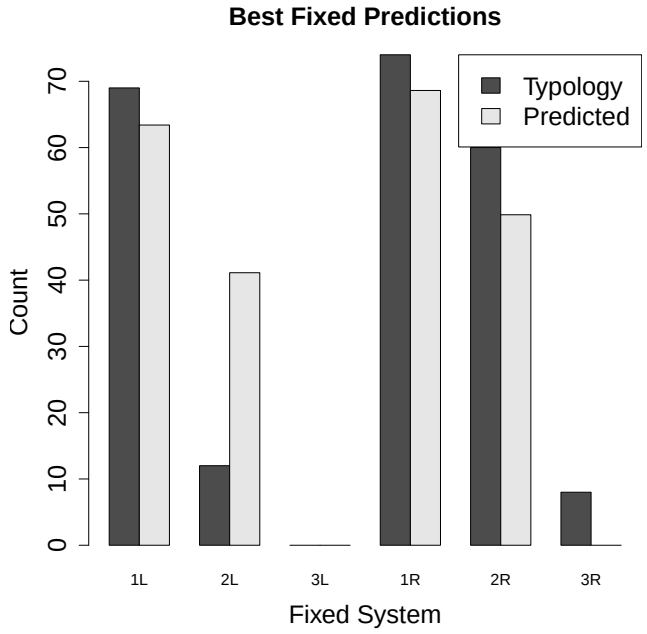


Figure 3.13: Typological predictions with best results from optimizations: Random stress reassignment with probability 0.55 and 415 iterations. NONINITSyll, typology from the Stress Pattern Database. 1L, 2L, 3L mean fixed stress on the first, second, and third syllable from the left; symmetrical for 1R, 2R, 3R on the right.

peninitial count than predicted. This prediction asymmetry points to the cause of the problem here: the Stress Pattern Database largely has typological frequency tracking with distance from the word edge (unlike WALS), but there is a large interaction—peninitial stress is much less common than initial stress, but penultimate stress is not much less common than final stress. The model must trade edge biases for nonfinality biases—interactions such as this are difficult to model.

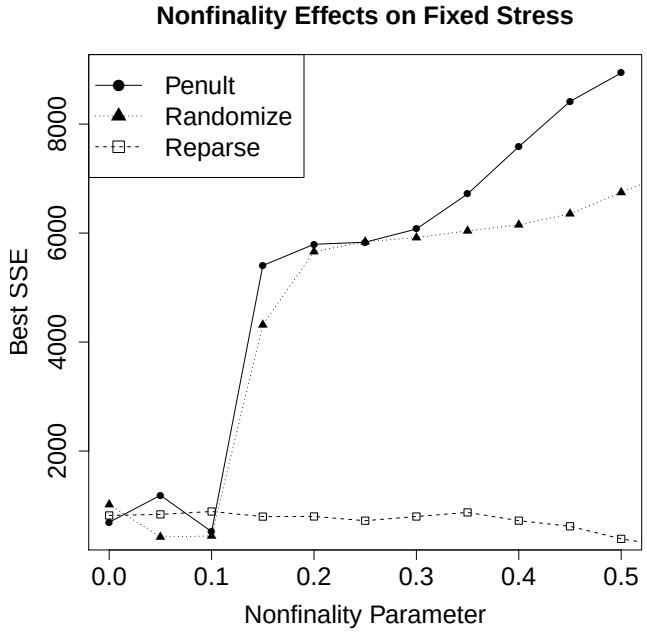


Figure 3.14: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. NONINITSyll not included, typology from the Stress Pattern Database. $\eta = 0.1$.

Results for the Stress Pattern Database when NONINITIALITY(Syllable) is excluded point to a solution to this problem of interaction. Figures 3.14 and 3.15 again show that there is no obvious evidence of a beneficial effect of the simulated nonfinality pressure. However, in the best results in Figure 3.16, we see that things have improved for the predictions on peninitial stress compared to the earlier Stress Pattern Database results shown in Figure 3.13. The reason is clear—we have removed a bias in *favor* of just this sort of stress, namely NONINITIALITY(Syllable). The re-

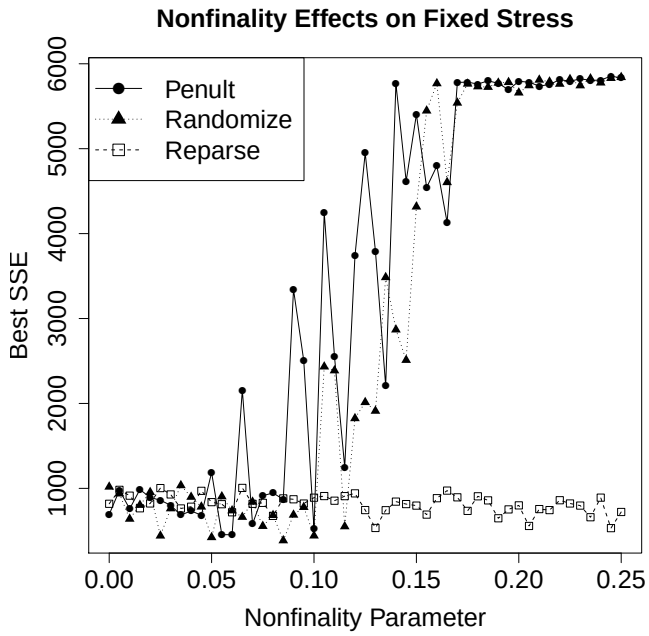


Figure 3.15: Probability optimization over a restricted range. NONINITSyll not included, typology from the Stress Pattern Database. $\eta = 0.1$.

maining constraints and learning bias substantially account for the existing typology, modulo remaining issues with prediction of postpeninitial and antepenultimate stress.

3.4.2.3 Summary

The results presented in this chapter show that there is clearly something to be gained from explicitly modeling substantive distinctions in linguistic patterns. Otherwise elusive, merely intuitive understandings rise to the point of prediction. The uncertain nature of the typology of fixed stress obscures the point somewhat, but it is clear that the qualitative predictions gained from modeling perceptual nonfinality are in line with the kinds of effects seen. Penultimate stress can be rendered more common or final stress less common as necessary. I have shown that two methods—simulating probabilistic misperception and altering assumptions on the constraint set—can affect results. Each of these is potentially advantageous, depending on the true typology and the way in which this set of assumptions is fused with existing learning biases.

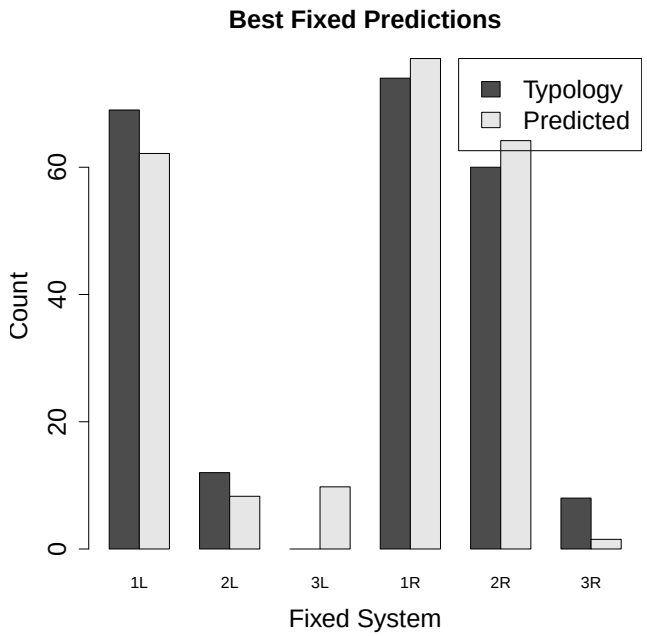


Figure 3.16: Typological predictions with best results from optimizations: Random stress reassignment with probability 0.085 and 2634 generations. NoninitSyll not included, typology from the Stress Pattern Database. 1L, 2L, 3L mean fixed stress on the first, second, and third syllable from the left; symmetrical for 1R, 2R, 3R on the right.

Further typological work can thus serve to disambiguate these approaches, with a ready account under either view.

3.4.3 Directional asymmetries in windows

In Chapter 2 I showed that some tendencies in the typology of stress windows can be accounted for by explanations based on relative learnability. In particular, I showed that learnability predicts that small windows should be more common than larger ones and that *some* categorical length limit should be observed. The latter fact follows even without a categorical limit imposed by the grammatical representation.

In that section, I mentioned a parallel with fixed stress: there is an overall bias towards the word edge, accounted for by learnability, but also a right/left asymmetry that is *not* accounted for. As seen in Figure 3.9, the frequency of windows of a given size at the right edge exceeds the frequency of windows of that size at the left edge. I propose that this asymmetry can be modeled using the same methods as fixed stress in §3.4.2—learnability combining with a nonfinality bias imposed on data.

Window type	Count	
Final two syllables	82	e.g. Malayalam (Asher and Kumari, 1997)
Final three syllables	38	e.g. Comanche (Smalley, 1953)
Initial two syllables	39	e.g. Yapese (Jensen et al., 1977)
Initial three syllables	1	e.g. Pirahã (Everett and Everett, 1984)

Table 3.9: Typological counts for window stress from StressTyp. Adapted from Kager (2012, ex. 22). Counts are collapsed across types of designated property and the position of default stress. Replication of Figure 2.12

A stress window language is one in which stress is required to fall within a certain number of syllables of the word edge, but the precise location of stress is determined by some other factor (some "designated property"). For example, stress might fall on whichever of the last two syllables of a word is heavy. This would be a two-syllable window at the right edge. When there is no designated property or it falls outside the window, stress is assigned in some default way.

A nonfinality perceptual bias could affect window systems in much the same way that it affects fixed stress. Windows at the left edge can require final stresses in exceptional cases, meaning that the part of the pattern not favored by learnability (i.e. edge alignment) is *also* not preferred by nonfinality. In contrast, any final stresses in a right-edge system *are* supported by their greater alignment. One could expect, then, that nonfinality introduces a pressure to probabilistically transform left-edge window systems into right-edge ones, yielding the typological tendency.

In my prior discussion of windows I assumed one of the simplest defaults possible: in the absence of other conditioning factors, stress falls on the syllable closest to the relevant edge (i.e. final in a right-edge system, initial in a left-edge system). This is not the ideal default to focus on for the purposes of examining edge asymmetries. The reason for this is that an edge default is very learnable—it has a highly reliable default position. This means that a left-edge window of this type will only very reluctantly be learned as a right-edge window (and *vice versa*). Another default position—one which is in fact more common—is more suitable to viewing how one type of edge orientation can be learned as another. This default type places stress—in the absence of a designated property—onto the syllable farthest *away* from the edge, while remaining within the window. For example, a two-syllable window at the right edge would have default stress on the penult. Windows of this type are shown in Figure 3.10. In simulations, I consider only these windows, not any fixed stress languages or any windows larger than the known attested instances This contrasts with Chapter 2. I do this in order to focus on the pressures exerted by nonfinality on window stress systems in isolation.

Results for windows mirror those for fixed stress. The error graph shown in Figure 3.17 exhibits the kind of local minimum we need to verify the effect of the nonfinality parameter. Again, it seems that the penult and random methods are more effective than the reparse one.

	$\sigma\sigma\sigma\sigma\underline{\sigma}$	$\sigma\sigma\sigma\underline{\sigma}\sigma$	$\sigma\sigma\underline{\sigma}\sigma\sigma$	$\sigma\underline{\sigma}\sigma\sigma\sigma$	$\underline{\sigma}\sigma\sigma\sigma\sigma$
"One-syllable"	$\sigma\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\sigma\acute{\sigma}$
Two-syllable	$\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\acute{\sigma}\sigma$	$\sigma\sigma\sigma\acute{\sigma}\sigma$	$\sigma\sigma\sigma\acute{\sigma}\sigma$	$\sigma\sigma\sigma\acute{\sigma}\sigma$
Three-syllable	$\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\acute{\sigma}\sigma$	$\sigma\sigma\acute{\sigma}\sigma\sigma$	$\sigma\sigma\acute{\sigma}\sigma\sigma$	$\sigma\sigma\acute{\sigma}\sigma\sigma$
Four-syllable	$\sigma\sigma\sigma\acute{\sigma}$	$\sigma\sigma\sigma\acute{\sigma}\sigma$	$\sigma\sigma\acute{\sigma}\sigma\sigma$	$\sigma\acute{\sigma}\sigma\sigma\sigma$	$\acute{\sigma}\sigma\sigma\sigma\sigma$

Table 3.10: Examples of window stress systems. If the designated property (underline) is within the window it is matched by surface stress. Otherwise default stress results. The default assumed here is stress on the syllable farthest from the edge, within the window.

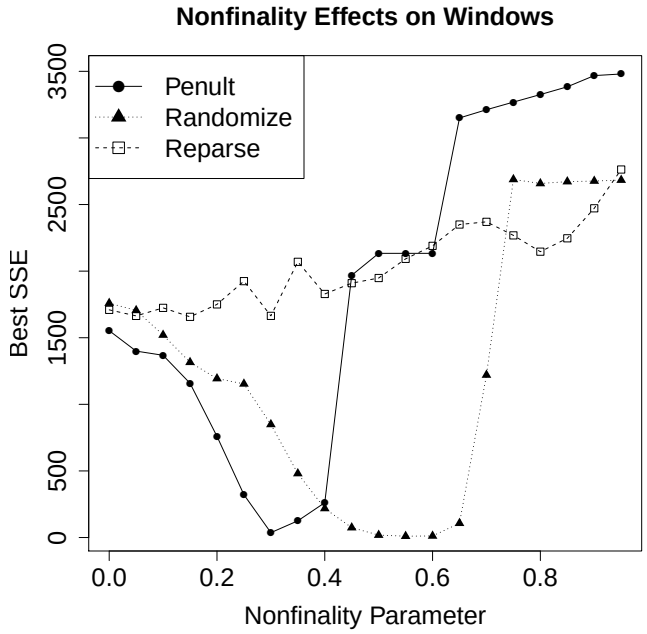


Figure 3.17: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. $\eta = 0.1$.

Results for the best-predicted typology are actually *more* successful that those seen for fixed stress. The model can find results that closely mirror the frequencies of two- and three-syllable windows at the left and right edges. A possible reason for this increased performance is that window stress frequency (over this range, at least) is monotonic with distance from word length (unlike the WALS data) and exhibits less interaction between this effect and word edge (unlike both fixed datasets). The best results obtained are shown in Figure 3.18.

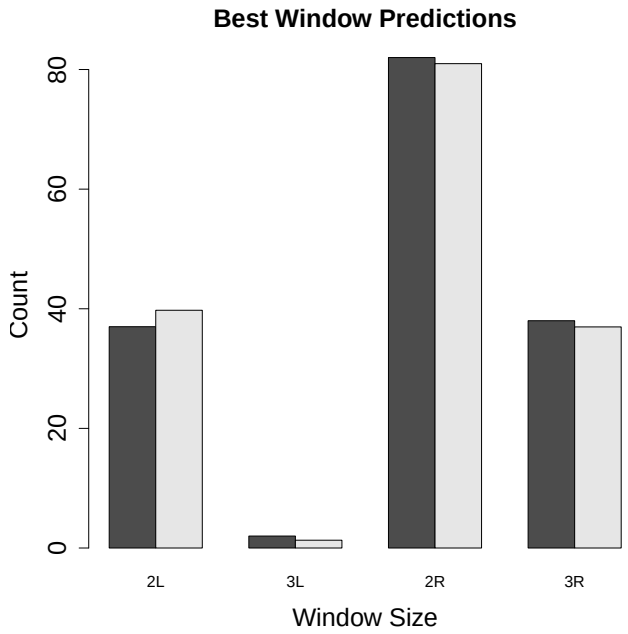


Figure 3.18: Typological predictions with best results from Figure 3.17: Reparse stress reassignment with probability 0.60 and 169 generations. 2L and 3L mean stress in windows of size two and three on the left; symmetrical for 2R and 3R on the right.

3.4.4 Iambs and trochees

In Chapter 1, I briefly discussed a broad, parameterized typology of iterative stress. This parameterization is demonstrated again in Table 3.11: iterative systems are divided according to whether they iterate left-to-right or right-to-left (with left-

to-right shown here), whether they use iambs or trochees, and whether they tolerate degenerate monosyllabic feet.

	No degenerate feet (Binary)	Degenerate feet (Nonbinary)
Trochaic	$(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma}\sigma)(\acute{\sigma})$
Iambic	$(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\sigma\acute{\sigma})\sigma$	$(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\sigma\acute{\sigma})(\acute{\sigma})$

Table 3.11: Parametric left-to-right patterns. Duplication of Table 1.5.

The broad frequencies for these types, derived from StressTyp, are shown in Table 3.12. One fact that is obvious at first glance is that iambic stress is less common than trochaic stress. This fact has been noted repeatedly (e.g. by Hayes, 1995). For the numbers given here, one could say that for any iambic pattern there is at least one trochaic pattern that is more common than it. In fact, this is weaker than what we see—for any given iambic pattern, all trochaic patterns are more common, other than right-to-left trochees with degenerate feet.

Iambs, being right-headed feet, serve to enforce stress later in the word than an equivalently-placed trochee. One view of them, then, is that iambic patterns have a tendency to require final stress. If the above accounts of nonfinality have merit for fixed and window stress, perhaps they can find application here as well—perhaps iambic patterns are uncommon due to the final stresses involved in many such patterns. In Table 3.13, I annotate each broad stress pattern with the degree to which it involves final stresses in "long words" (viz. words longer than two syllables). Some patterns mandate final stress, others ban it, and still others require it for exhaustivity.

Strikingly, no pattern annotated as "always" having final stress is common. All three of these are iambic systems, as suspected. The one iambic system with any frequency is one in which final stress occurs only in even parity words. In addition, this system is a perfect grid system, shown in Chapter 1 to be potentially advantageous in learning.

Foot type	Direction	Degenerate feet?	Count
Trochees	Left-to-right	no	33
		yes	22
	Right-to-left	no	34
		yes	4
Iambs	Left-to-right	no	13
		yes	3
	Right-to-left	no	2
		yes	3

Table 3.12: Parametric iterative stress in StressTyp. *Degenerate feet?* indicates whether all feet are binary. That is, "no" indicates that degenerate feet are not permitted. Duplication of Table 1.6.

Foot type	Direction	Degenerate feet?	Perfect grid?	Final?	Count
Trochees	Left-to-Right	no	no (right lapse)	never	33
		yes	yes	sometimes	22
	Right-to-Left	no	yes	never	34
		yes	no (left clash)	never	4
Iambs	Left-to-Right	no	yes	sometimes	13
		yes	no (right clash)	always	3
	Right-to-Left	no	no (left lapse)	always	2
		yes	yes	always	3

Table 3.13: Iterative stress typology, annotated with status as perfect grids and degree of final stress required.

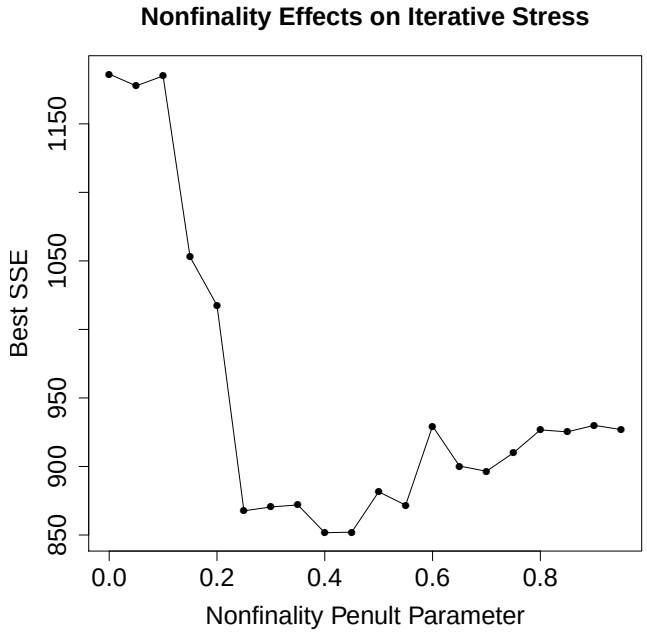


Figure 3.19: Best result of optimizing SSE across iteration counts for a range of probabilities and nonfinality simulation methods. Only penult stress assignment considered. $\eta = 0.1$.

Figure 3.19 shows the result of optimization over the iterative stress typology, focusing on penult stress assignment only. In these simulations, PARSE(Syllable) is of use to guarantee relatively exhaustive parsing. We see that here too there is an apparent effect of the nonfinality parameter—including this effect allows a potential reduction in the typological prediction error.

Figure 3.20 shows the best result for typological prediction. The fit is far from perfect. Many languages are given a predicted zero frequency, even if they are actually well-attested. This is the perfect grid effect run rampant—the languages with zero frequencies are just the non-binary ones (tolerating degenerate feet—labeled without a B in the figure). These languages offer less reliable placement of stress with respect to constraint violations, and are therefore dispreferred on learning grounds to their counterparts with no degenerate feet. This means that, in the limit, a small bias

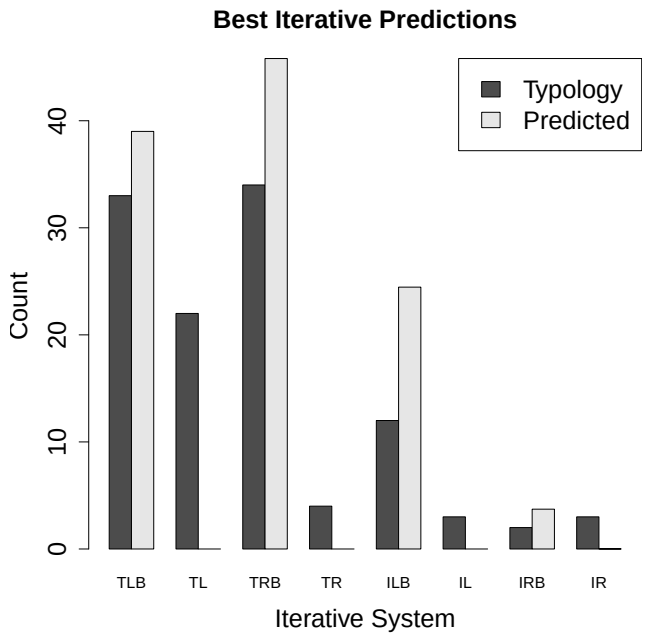


Figure 3.20: Typological predictions with best results from Figure 3.19: Penult stress reassignment with probability 0.40 and 169 generations. T/I indicate a trochaic/iambic parse. L/R indicate a parse from left-to-right/right-to-left. B indicates a pattern that is strictly binary (that is, does not tolerate degenerate feet). See Table 3.13.

against degenerate feet ends up resulting in a typology with only systems without such feet.

However, of the languages predicted, the order is correct: iambs are less frequent than trochees overall, right-to-left trochees are more common than left-to-right (marginally), and left-to-right iambs are more common than right-to-left. Thus this model successfully predicts that iambic systems should be dispreferred. The component of the model of particular interest—nonfinality—exhibits the kind of pressure desired, pushing for trochaic parses. It is of less concern for this particular set of results how frequencies are predicted between systems with and without degenerate feet. The important focus here is on an overall push for trochees at the expense of iambs, as seen in Figure 3.21.

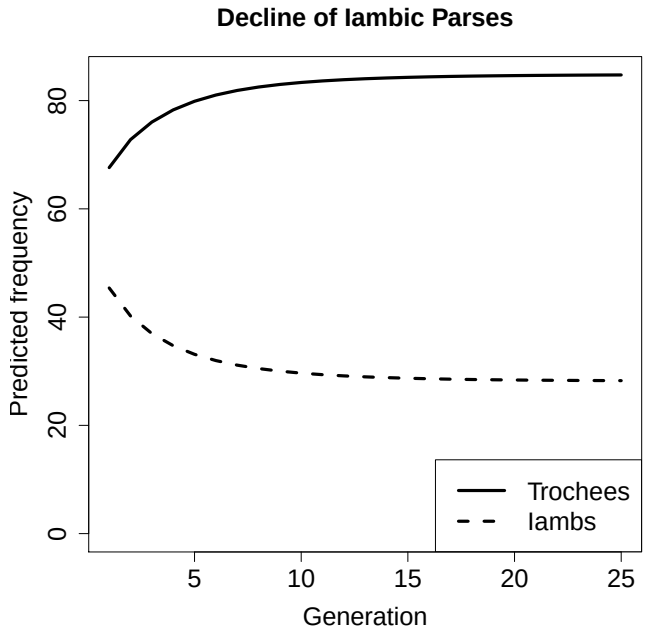


Figure 3.21: Growing typological dominance of trochees over generations, using best results from Figure 3.19. Probability is consolidated onto trochaic parses as the number of generations increases, resulting in predictions as shown in Figure 3.20.

The reduction in error, coupled with this breakdown of results, shows that the nonfinality pressure does useful work despite the problems posed by learning iterative

systems with degenerate feet. This confound points to a need to better understand how different learning forces should be balanced against one another, but it need not invalidate them.

3.5 Conclusion

In this chapter, I have demonstrated that explicit models of learning allow similarly explicit tests of extragrammatical pressures on linguistic typology. In Chapters 1 and 2 I showed that learning models can expose predictions made by grammatical assumptions for probabilistic typology. These predictions do not likely exhaust the set of useful explanations for the relative attestation of linguistic patterns. Biases emerging from outside of grammar are expected to have their part as well. I showed that assumptions about these kinds of pressures can be integrated with a learning model to make combined predictions about typology.

The specific example used was nonfinality. A hypothesized perceptual issue with final stress was modeled in several ways as probabilistic misperception of data with final stress. This perceptual simulation was shown to give improvements with respect to models of typological frequency for fixed stress, windows, and general iterative patterns. Future work using this methodology could proceed by introducing probabilistic variation in data of other types.

CHAPTER 4

CONCLUSIONS

4.1 Contributions

4.1.1 Overview

In this dissertation, I show a connection between the typological frequency of stress patterns and their relative learnability. I demonstrate that grammatical assumptions, coupled with a learning algorithm, naturally lead to distinctions among possible languages—some are learned quickly, others more slowly. I give explicit examples of frequency prediction, primarily within the domain of stress typology.

This type of approach is significant in that it opens up a richer typology for analysis. The researcher does not need to be concerned only with distinctions between possible and impossible languages, being freed to explain more nuanced distinctions even with the familiar mechanisms of typical generative linguistic theory. This has two primary advantages. First, it allows an explanation of a linguistic fact that is known but seldom modeled explicitly: not all patterns are equally common. This is positive in that the field gains greater empirical coverage. Second, the analyst is not tied to making important empirical judgments about the difference between impossible languages and possible but unattested languages—a distinction that is fraught at best.

This work shows how this sort of enterprise can proceed, giving explicit examples. Future work can follow to build an even more unified view of typological frequency in stress, and to extend these methods to larger linguistic domains.

4.1.2 Fixed stress and alternation

In Chapter 1, I showed that simple n-gram constraints create learning biases conditioning both fixed stress and types of alternation. Fixed stress languages with stress close to a word edge are more quickly learned than languages with stress farther away from the edge. That is, initial stress is learned more readily than peninitial, which is learned more readily than postpeninitial, etc. This bias emerges because short edge distances exhibit less variability across word lengths. Small words show more of the same n-grams as large ones, so learning in any word length tends to push to the same constrained set of hypotheses. As the distance to the word edge increases, different word lengths inform on different hypotheses, slowing learning.

This same logic was echoed in my discussion of perfect grids. Full alternation of stress gives a situation in which stress determination is highly local. Reflected in the n-gram constraints, this means that violations of the constraints are highly reliable. Again, word lengths will be quite similar in the information they contribute toward learning a language, aiding learning.

Both of these cases can be viewed as an explanation of isolated pieces of stress typology. Stress toward word edges is, broadly, more common than stress farther away. Alternation patterns are generally much more "simple" and local than random chance. Thus I show a correspondence between the bias projected by simplified assumptions on a representation and the typology, providing support for the idea that learning biases are useful for explaining probabilistic typology.

4.1.3 Primary stress correlations and window stress

In Chapter 2, I showed biases relating to a constraint set constructed from alignment and rhythmic constraints. I show that the kinds of biases found in Chapter 1 resurface for other types of typological tendencies. The model shown is biased towards languages in which primary stress is in a relatively fixed location. This follows

because constraints pertaining to primary stress will then be less variably violated across word lengths, leading to more reliable and fast learning.

The first important consequence of this type of bias is for the placement of primary stress in an iterative stress parse. Parses in which secondary stress is left-to-right will be learned better if their primary stress is on the leftmost stressed syllable. Likewise, by symmetry, for right-to-left parses. The only departure from this is for so-called bidirectional systems, in which one foot is reliably isolated at an end of a word, regardless of overall directionality. In such languages, the bias is for that isolated foot to bear primary stress. These biases emerge simply because they result in more reliable representations of primary stress. They have the important consequence of predicting—in an iterated learning model—that languages obeying this correlation between directionality and primary stress position should be more common. This predicted tendency is in fact evidenced typologically.

The reliability effect found by consistent placement of primary stress has other consequences. I show that the model exhibits biases for stress in stress window systems to fall close to a word edge. This mirrors not only the results on primary stress correlation but also the fixed stress results of Chapter 1. Stress windows are predicted to be small overall, with frequency increasing as the size of the window decreases. This prediction is borne out by typological study: two-syllable windows are more common than three-syllable ones. A further contribution of this section, in addition to its approach to overall window size, is a discussion of the apparent categorical cutoff at four syllables. No four-syllable windows are robustly attested. This could be attributed to the categorical means available to learners for representing language, but another approach is the one shown. Four-syllable windows are simply an accidental gap, but their absence is somewhat expected—they are biased against to such a degree that it should not be surprising to only see windows of larger size.

4.1.4 Nonfinality simulation

In Chapter 3, I confront the issue of biases emerging from outside the grammar itself. If grammatical assumptions are symmetrical, as assumed in Chapters 1 and 2, something is necessarily required to break this symmetry. This is because the typology of stress—and linguistic patterns generally—does not obey formal symmetries overall. In stress this is most obvious in the instance of left/right asymmetries: a pattern is unlikely to be equally attested as a mirror-image pattern with all strings flipped left-to-right. In this chapter, I specifically address a putative nonfinality bias, with the goal of breaking this symmetry. This bias creates a representational or perceptual problem with final stress that does not exist with initial stress, causing patterns oriented to the left edge to have distinct predicted frequencies from ones oriented to the right.

I show that this kind of modeled nonfinality is potentially useful in capturing otherwise unexplained typological tendencies. Modeling nonfinality as a probabilistic tendency for final stress data to be misinterpreted as something different from final stress, I show that several typological skews in stress potentially follow from nonfinality.

The first such skew is found in stress windows. The discussion of windows in Chapter 2 ignored the fact that windows are more common at the right edge than the left. I show that this distinction follows from the addition of biases pushing stress off the right edge. Simply, there is a perceptual reason for stress windows at the right edge that is absent for windows at the left. It therefore follows that windows should be more common, and more complex, at the right edge.

The second tendency is found in fixed stress, discussed initially in Chapter 1. Fixed stress systems, similar to windows, are asymmetrically attested. I show that the same kind of "push" off the right edge is useful for understanding why penultimate stress

might be comparatively abundant and/or final stress comparatively absent, taken in contrast with symmetrical patterns at the left edge.

Finally, this perceptual bias can potentially illuminate a difference between iambic and trochaic parsing. Trochees (left-headed feet) tend to create final stressless syllables, while iambs (right-headed feet) create final stresses. Only the latter opposes the nonfinality bias, and it is this sort of pattern that is typologically underattested.

4.2 Review of methodology

In this dissertation, I used a consistent set of theoretical tools in order to probe predictions of models for probabilistic tendencies in typology. In this section, I will explicitly set out the components needed (or useful) in an investigation of this type. This section is intended as a guide to replications or extensions of work of the type shown in this thesis.

4.2.1 Grammatical assumptions

Predicted probabilistic typologies derived from learning require exact understandings of the assumed grammatical model, just as categorical predictions do. In this regard, the methods described in this dissertation do not differ from typical generative phonology. In this thesis, the grammars used are always Maximum Entropy (Goldwater and Johnson, 2003) grammars. MaxEnt grammars are weighted; this fact allows results such as those on stress windows in Chapters 2 and 3. The constraint sets assumed throughout are either the n-gram constraints of Chapter 1 or the augmented gradient alignment constraint set used elsewhere.

In the work presented here, the set of candidates assumed is perhaps more limited than typical typological work. For example, word lengths are constrained to be between two and eight syllables. When analysis involves computational simulation,

the examination of such assumptions proceeds differently from non-computational work, but with the same goals.

4.2.2 Distributional assumptions

In simulation work, the analyst must make modeling assumptions beyond grammar-internal ones. Various other forces come into play, manipulating the nature of the data as it is transmitted from teachers to learners.

In an OT-like framework, the first consideration is the distribution over members of the languages being learned. A learner is not likely to encounter data in a uniform fashion across all types of datum. Instead, there will be concentrations of probability on certain parts of the linguistic system, principled by some concerns that are essentially external to grammar *per se*. The analyst must decide what types of distribution should be reflected in the model, and in what way.

In this dissertation, I generally model the probability of word lengths as an exponentially decreasing function. This choice is not motivated by anything internal to assumptions about how grammar works. Instead, it is driven by observations that languages do in fact seem to pattern their word lengths in this way. An explanation of this bias is assumed to lie elsewhere. A less informed decision is made about the distribution over positions of "designated properties." These properties are assumed to arise uniformly throughout a word, perhaps in opposition to actual (but unknown) cross-linguistic tendencies.

This notion of distributional assumptions can be extended away from thoughts about probabilities over forms to assumptions about the qualities of the transmission channel between teacher and learner. A learning model necessarily makes assumptions on whether the learner receives data veridically—and how data is received when this transmission falters.

In Chapter 3 I utilize this component of assumptions to investigate models of a perceptual nonfinality bias. I show that this component of assumptions can have qualitative impact on the nature of predictions made by a typological model incorporating learning.

4.2.3 Learning assumptions

A model of learning must, of course, make assumptions on learning. An assumed learning algorithm should ideally be motivated by outside considerations: simplicity, use elsewhere, psychological plausibility, etc. Learning bias results should be explored for a range of reasonable parameter values for the learning algorithm, establishing the relative fragility or robustness of the results. A fragile result is not necessarily wrong, but calls for further work.

In this thesis I used a version of the SGA (Jäger, 2007) for MaxEnt grammar. A range of values of the learning rate parameter were checked, but results throughout use a consistent set of assumed values.

A related set of assumptions couples the learning algorithm to typology. The analyst must ask what model of typological shift counts as sufficiently true to life in order to draw inferences. This is done in particular by assuming particular network structures to an iterated learning method. In this dissertation I assume a simple "chain" form of iterated learning.

4.2.4 Learning results

In this work, I show two methods of revealing the biases in a learning model. Both are valid for certain purposes, but have potentially distinct goals. In the first, I compare the residual error of a learner with respect to its teacher after some amount of learning. This comparison illuminates the raw biases that exist in the process of learning—that is, this can answer the question of exactly how learning works out differently for learners of different languages.

This focus on error contrasts with approaches focused on predicted typologies. If we wish to model typological differences, and not just error differences, we must explicitly model typological emergence in some way. This is because learning biases need not be transparently reflected in the typological prediction of a given model of e.g. language change (Rafferty et al., 2011). This is where models of iterated learning, for example, come into the picture. The analyst derives predictions about the way in which one language type changes into another, and thence derives predicted frequencies over language types.

This process is not assumption-free. The analyst must ask if change is well-modeled by a probabilistic change between categorical language types, or if languages themselves are probabilistic. I have used both methods, but principally present the former in this dissertation. It is additionally important to ask "where" in change we should evaluate results. Should this be at convergence, or some earlier point? When this question arises, I have taken the stance that it is best to always be comparing models. When there is a comparison, each can be given its best chance to succeed (i.e. its moment of best performance can be chosen), and any choice between the models reflects an effort to give each hypothesis its best fighting chance.

4.2.5 Overall

This emphasis on comparison guides the overall structure of the learning work presented here. Linguistic typology is an uncertain thing; probabilistic typology is more so. Languages types could easily have been missed by analysts or have not arisen by accident of contingent human history. The typology that we do have could be shaped by that same history—wars destroying a language group exhibiting a certain pattern, technological advantage leading to the abundance of another pattern, and so on. Analysts might have mischaracterized individual languages, ultimately adding them to the wrong frequency count within a probabilistic typology. On the opposite

end, modeling typology is also uncertain. All the assumptions above are difficult to disentangle. It would be a rare thing indeed to be certain of all conditions save the one of interest.

These issues, and more, should not lead us to abandon modeling probabilistic typology. Introducing frequency counts does not make our models more subject to these concerns, and could even help. Where these concerns lead me, instead, is to view typological modeling as a set of comparisons. Language types should be counted coarsely, eliminating many of the small variants that obscure typological accounts. Having done this, we can be more certain of the typology of interest. Further, we can be more satisfied with our typological modeling assumptions; small changes in assumptions might make a big difference for individual linguistic peculiarities in a way they would not make a difference for these coarse-grained predictions.

4.3 Future directions

4.3.1 Other stress tendencies

This dissertation by no means exhausts the list of typological tendencies in stress. This is particularly true with biases that might have extragrammatical conditioning factors. For example, the Iambic-Trochaic Law (Hayes, 1985) establishes a connection between the correlates of stress and the headedness of metrical grouping. Stress distinctions based on intensity tend to be left-headed (i.e. trochaic), while distinctions based on duration tend to be right-headed (i.e. iambic). One might view this typological generalization as emerging from probabilistic perceptual pressures on learners, just as nonfinality was discussed in Chapter 3.

Another example is an apparent typological avoidance of clash between primary and secondary stress. Two adjacent stresses are apparently more marked if one is primary (Kager, 2001), evidenced in particular languages such as English (Pater, 2000). A probabilistic approach to misperception of stresses could help ground this

bias without necessarily having a typology-wide markedness hierarchy. This would be useful in particular within my approach due to overprediction (due to reliability) of just such languages causing primary/secondary clash (Chapter 2).

4.3.2 Morphological patterns

Morphological exponence can be thought of as operations transducing one form into another. The most common such operation is affixation (i.e. string concatenation), but many others are possible. Reduplication, templatic truncation, ablaut, and subtractive truncation are just a few. In other work I propose a model of morphology in which these operations are relatively free in their range of possible forms (Staubs, 2011). In that work I note that such operations are not a full explanation without some way of understanding why affixation is so common—or broadly, why there are frequency differences between different types of morphological exponence. This dissertation suggests a way of understanding such frequency distinctions as emergent from the relative learnability offered by different transductions.

4.3.3 Feature economy and simplicity

Extending initial results by Pater and Moreton (2012), Pater and Staubs (2013) find that learning in models similar to the ones developed in this dissertation can lead to predictions of relative feature economy. Future work could concern more of the ways in which phonologies are organized around repeated structural symmetries and how learning can explain the frequency of these regularities.

APPENDIX

GENETIC (IM)BALANCE IN STRESSTYP

In Chapter 1 I mention the issue of balance in the typological databases used for my generalizations about bias. These databases are not balanced for genetic affiliation or area of the world, resulting in oversampling of particular language families and geographic regions. In this appendix I present evidence that the numbers derived from the typology are relatively representative of the true typology. To accomplish this, I give the numbers derived from StressTyp (Goedemans, 2010) under two types of random resampling.

A.1 Demonstrating imbalance

The first question to address is whether StressTyp (Goedemans et al., 1996b) is, in fact, genetically imbalanced. The answer to this is a decisive "yes." To show this, we can use metadata already present in StressTyp. StressTyp contains two columns of data reflecting genetic affiliation: "Dialect of" and "Genetic Info." The genetic information contains a list of genetic classifications such as this one for Aguacateco: *Mayan, Quichean-Mamean, Greater Mamean, Ixilan.* Thus each lists contains a genetic classification in descending order of size. Here I call each level of this list a *genetic depth* and use this number in studying genetic balance. There are a maximum of 14 different genetic depths, disregarding the level of the language itself (= depth 15) and dialect (= depth 16), which I exclude. Any language with a total depth lower than 14 is assumed to duplicate its final depth across all remaining classification. That is, a full classification for Aguacateco contains additional layers of Ixilan

before the language level. Figure A.1 shows the relative diversity of each depth in the StressTyp descriptions. The classifications in StressTyp are, of course, a summary of diverse opinions on linguistic phylogenetics (and may conflate area and genetics at points), but serve as a useful approximation.

Once StressTyp can be parsed into genetic depths, we can immediately detect the imbalance of the dataset. At the lowest genetic depth (= 1), there are 78 different classifications. Of these classifications, only 43 have more than one language in them. There are 510 languages in the typology, but the top five classifications make up 285 (56%) of these. The top five classifications are Austronesian (117 languages), Australian (66 languages), Indo-European (57 languages), Afro-Asiatic (24 languages), and Trans-New Guinea (21 languages). Thus at the largest level of classification, StressTyp has very many of very few classifications—a lack of balance. As genetic depth increases, languages are spread across more classifications, moving into the tail of the distribution. However, there remains a notable concentration around a few classes even at the deepest level (Figure A.2).

A.2 Biased resampling

StressTyp's lack of balance is not enough on its own for worry. This imbalance need be such that the counts presented in this dissertation misrepresent the shape of the typology. One way to establish whether or not this is true is to resample the typology from the StressTyp genetic classifications discussed above. If many resamplings result in similar relationships between language classes, these relationships can be considered fairly robust.

We know that StressTyp is imbalanced. One way to resample a pseudo-StressTyp, then, should mirror this imbalance—but impartially. We know that at depth 1, there are 117 languages in the largest class (Austronesian). A biased resampling of the data would have the same number of samples in its largest class, but the class is chosen

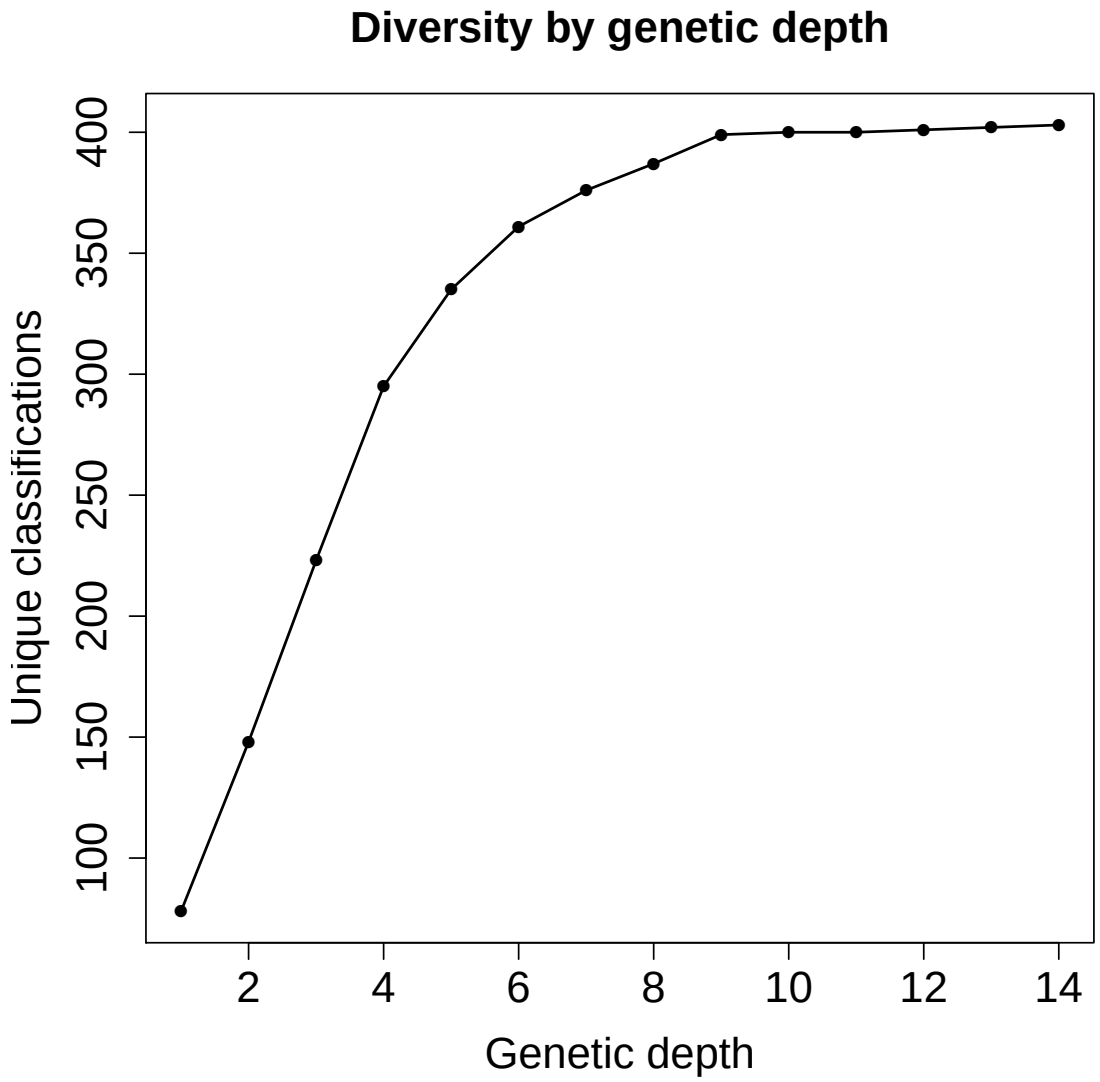


Figure A.1: Diversity at each genetic depth in StressTyp. A count of how many classification distinctions are made at each genetic depth.

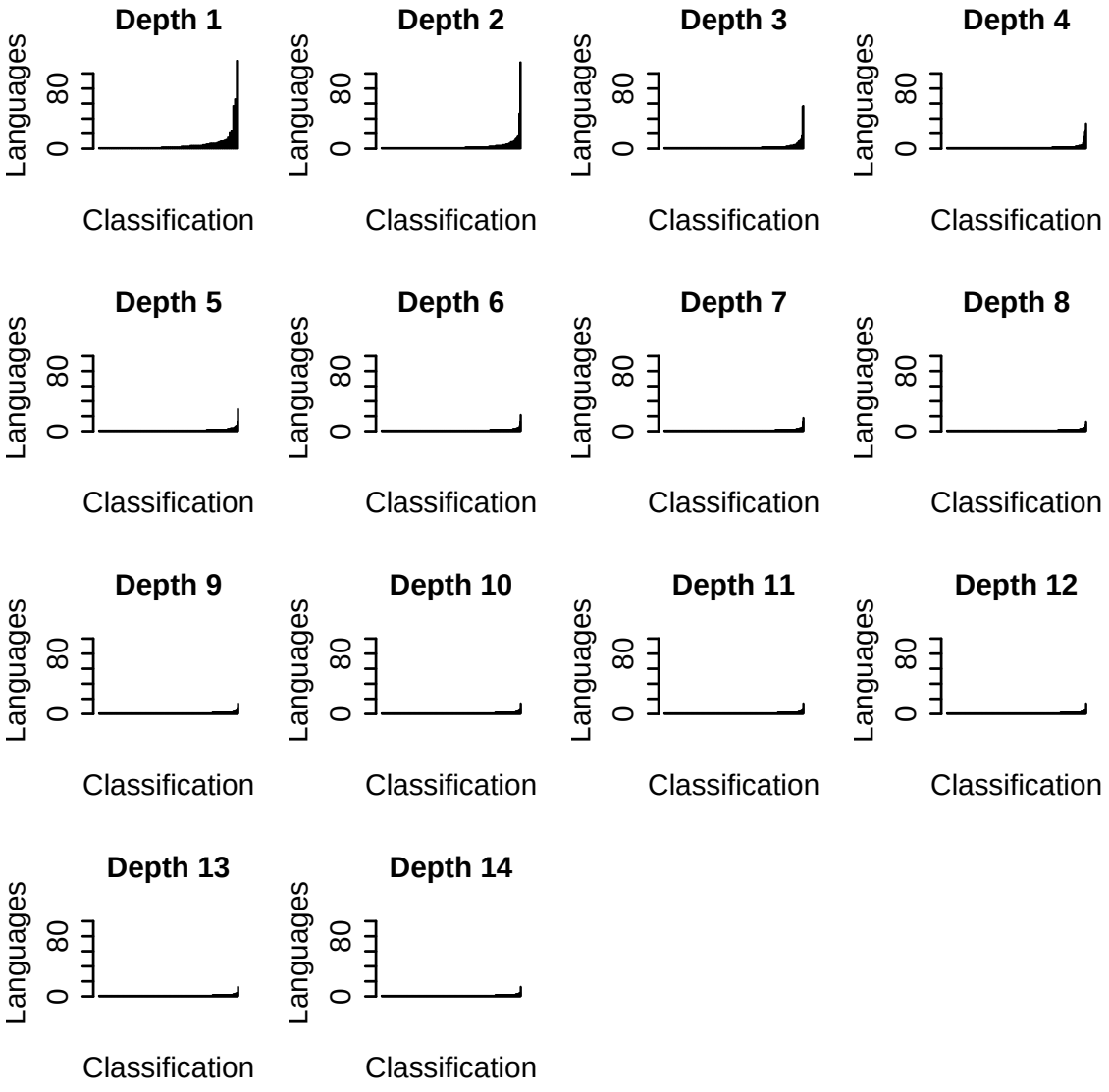


Figure A.2: Number of languages in each classification found in each depth. As depth increases, the number of classifications increases as well, so the number of languages decreases. Some classifications always have more languages than others.

uniformly at random from the options at that depth. One sample might yield Mayan as the "top" class. There are not 117 Mayan languages in StressTyp, and therefore the sample of languages within this class is necessarily with replacement. Sampling is uniform within a class. The goal of this kind of resampling is to establish whether the magnitude of imbalances within StressTyp, placed differently, is sufficient to shift results.

Figure A.3 shows results of this type for counting stress. Counting stress never accounts for more than a handful of languages in a sample, reinforcing evidence for the bias against these languages claimed previously. In fact, no sample has as *many* counting languages as in StressTyp—they are seemingly *oversampled* in the database, not undersampled.

For window stress (Figure A.4), two claims were made in the analyses. First, two-syllable windows are more common than three-syllable windows on the same word edge. This is largely the case in resampling, though for low genetic depths there is some confusion for right-edge windows. Second, right-edge windows are more common than left-edge ones. Again, this summarizes most of the typology, though with some over-sampling at low genetic depth.

Finally, in StressTyp fixed stress follows this order of attestation: initial stress, penult stress, final stress, peninitial stress, antepenultimate stress, postpeninitial stress. This order is observed at all genetic depths in resampling (Figure A.5).

A.3 Uniform resampling

Another approach to resampling disregards the type of imbalance found in the StressTyp counts. Instead, classifications are sampled uniformly at random with replacement at a given genetic depth, and a language is sampled from that classification. Results for this type of sampling appear largely as "smoothed" versions of results from the previous section.

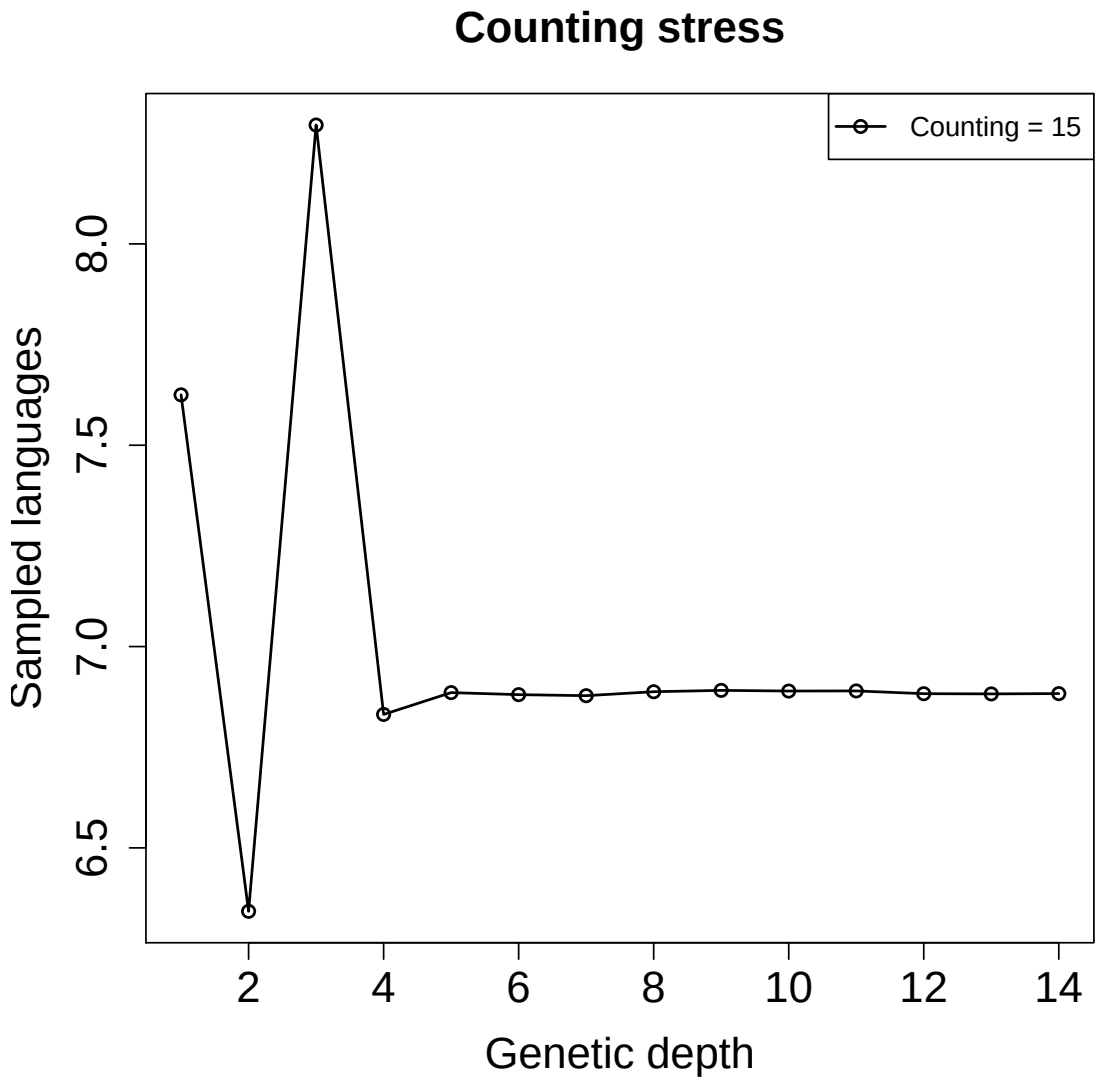


Figure A.3: Number of counting stress languages found in 50,000 resamples of the data, biased as in StressTyp. Legend shows observed value.

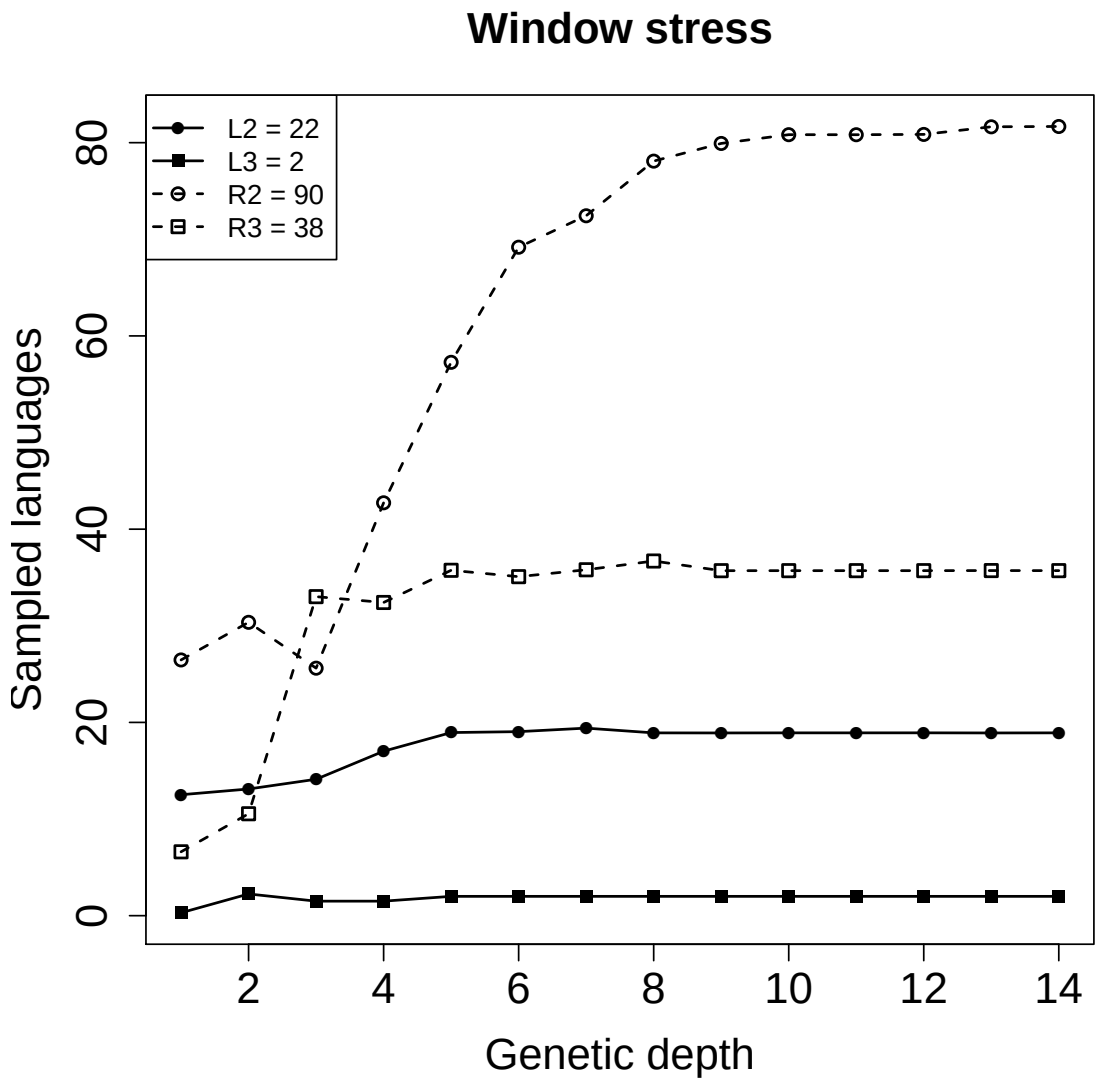


Figure A.4: Number of window stress languages found in 50,000 resamples of the data, biased as in StressTyp. L2, L3 are two- and three-syllable windows at the left edge. R2 and R3 are two- and three-syllable windows at the right edge. Legend shows observed values.

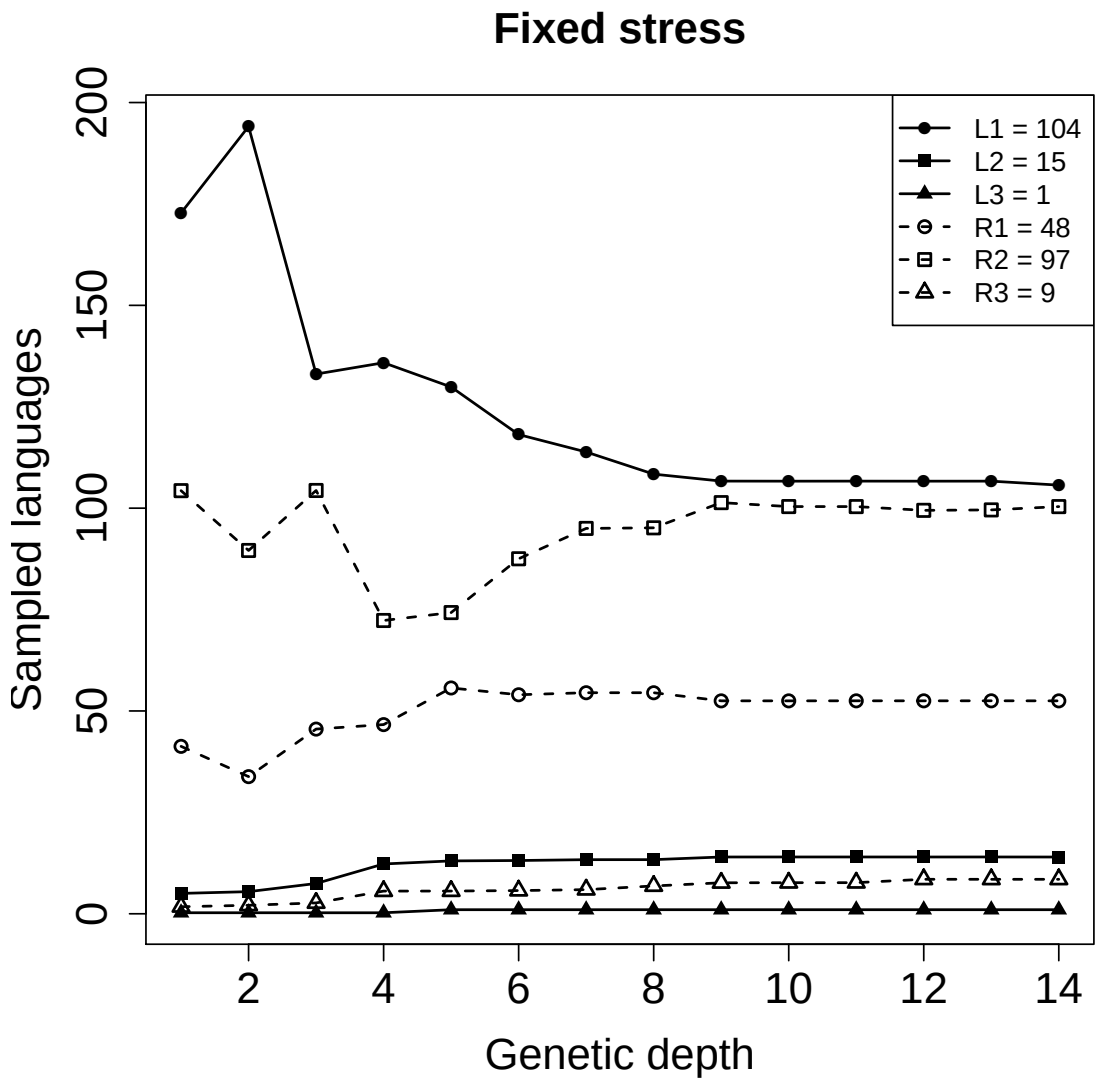


Figure A.5: Number of fixed stress languages found in 50,000 resamples of the data, biased as in StressTyp. L1, L2, L3 are initial, peninitial, and postpeninitial stress. R1, R2, and R3 are final, penultimate, and antepenultimate stress. Legend shows observed values.

Results for counting stress appear much as before, only now including the value 15 in their range (Figure A.6). Window stress again reflects some possible impacts of oversampling at low depths (Figure A.7), with overall agreement with claimed tendencies. Fixed stress is in great accord with the StressTyp numbers overall (Figure A.8). The only departure in fixed stress is the close approach between final and penultimate stress at low genetic depths. This is actually an encouraging result—StressTyp shows higher numbers of penultimate stress than final stress, while other typologies seemingly do not (see Chapter 3). These resampling results suggest that this confusion might be explainable if StressTyp oversamples low genetic depth classifications with a high propensity for penultimate stress.

A.4 Discussion

The analysis in this appendix shows that StressTyp is indeed not balanced with respect to genetic classifications. I provide evidence to suggest that similar sorts of typologies, with different incidental sampling biases, would obtain similar results for the large-scale numerical biases pursued in this dissertation. These results therefore serve to partially assuage worries that these biases are merely epiphenomena created by the sampling bias from analysts.

I have only examined StressTyp (Goedemans et al., 1996b), not the Stress Pattern Database (Heinz, 2007). This is despite using information from Heinz's database in parts of this dissertation. This is not due to any inherent difference in the databases' content—the Heinz database could be analyzed in a parallel manner. However, the majority of the typological claims made in this dissertation are supported in StressTyp, and thus study of its balance is most important. This is least true in the typology of fixed stress. As noted above, study of Heinz's database would be useful here, to determine if disagreements on final and penultimate stress frequency are due to different sampling biases.

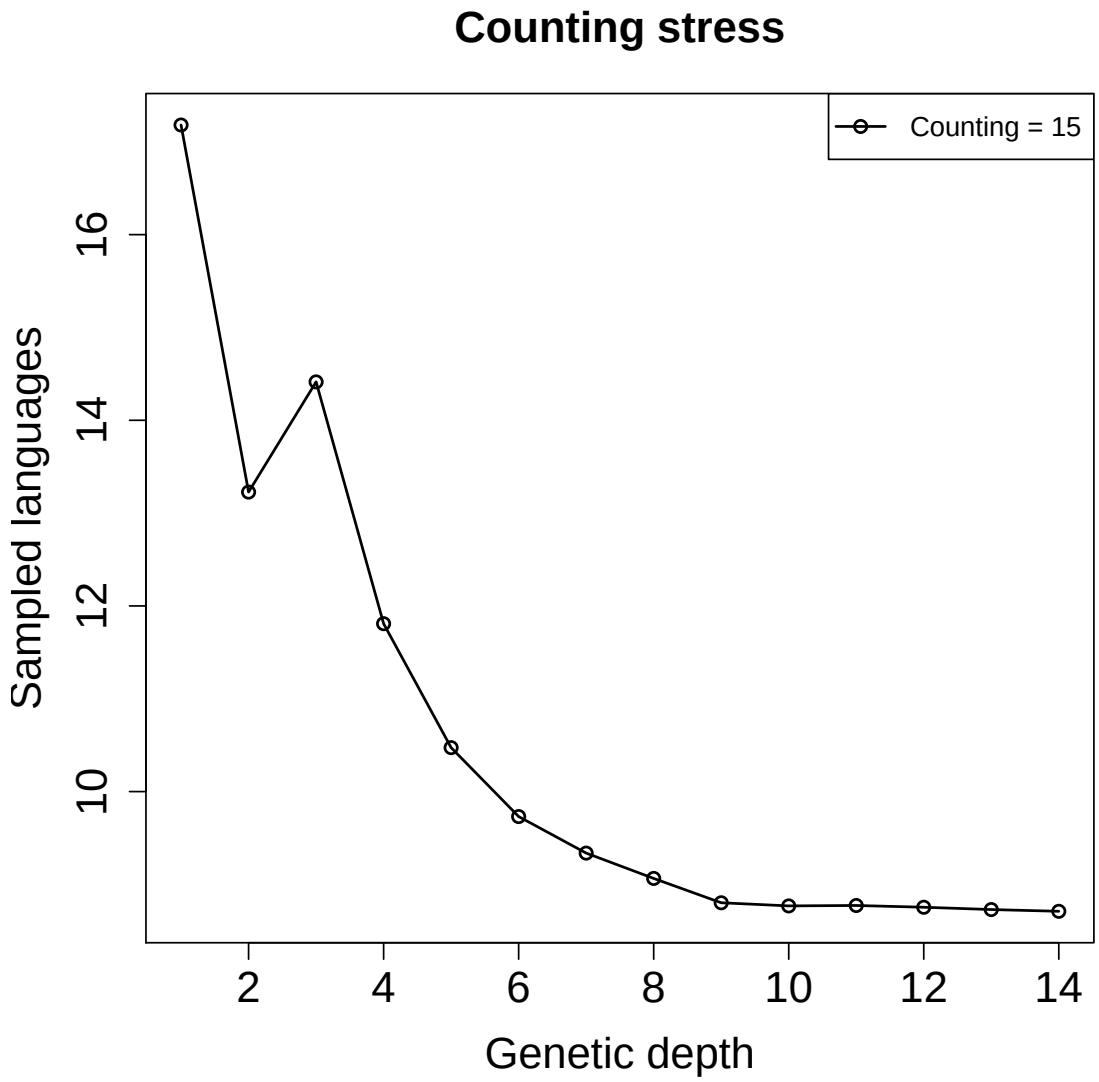


Figure A.6: Number of counting stress languages found in 50,000 resamples of the data, uniformly sampled within a genetic depth. Legend shows observed value.

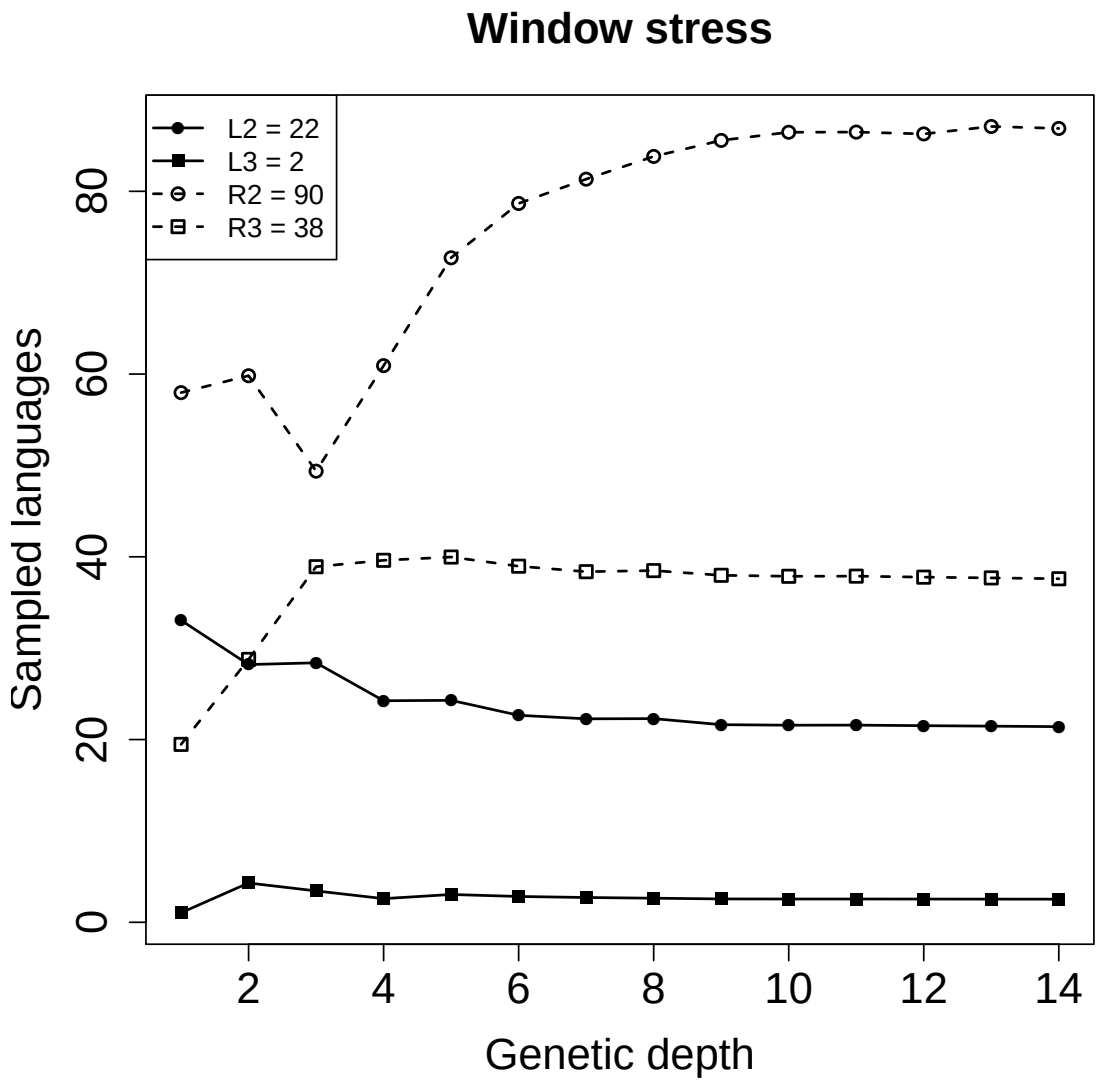


Figure A.7: Number of window stress languages found in 50,000 resamples of the data, uniformly sampled within a genetic depth. L2, L3 are two- and three-syllable windows at the left edge. R2 and R3 are two- and three-syllable windows at the right edge. Legend shows observed values.

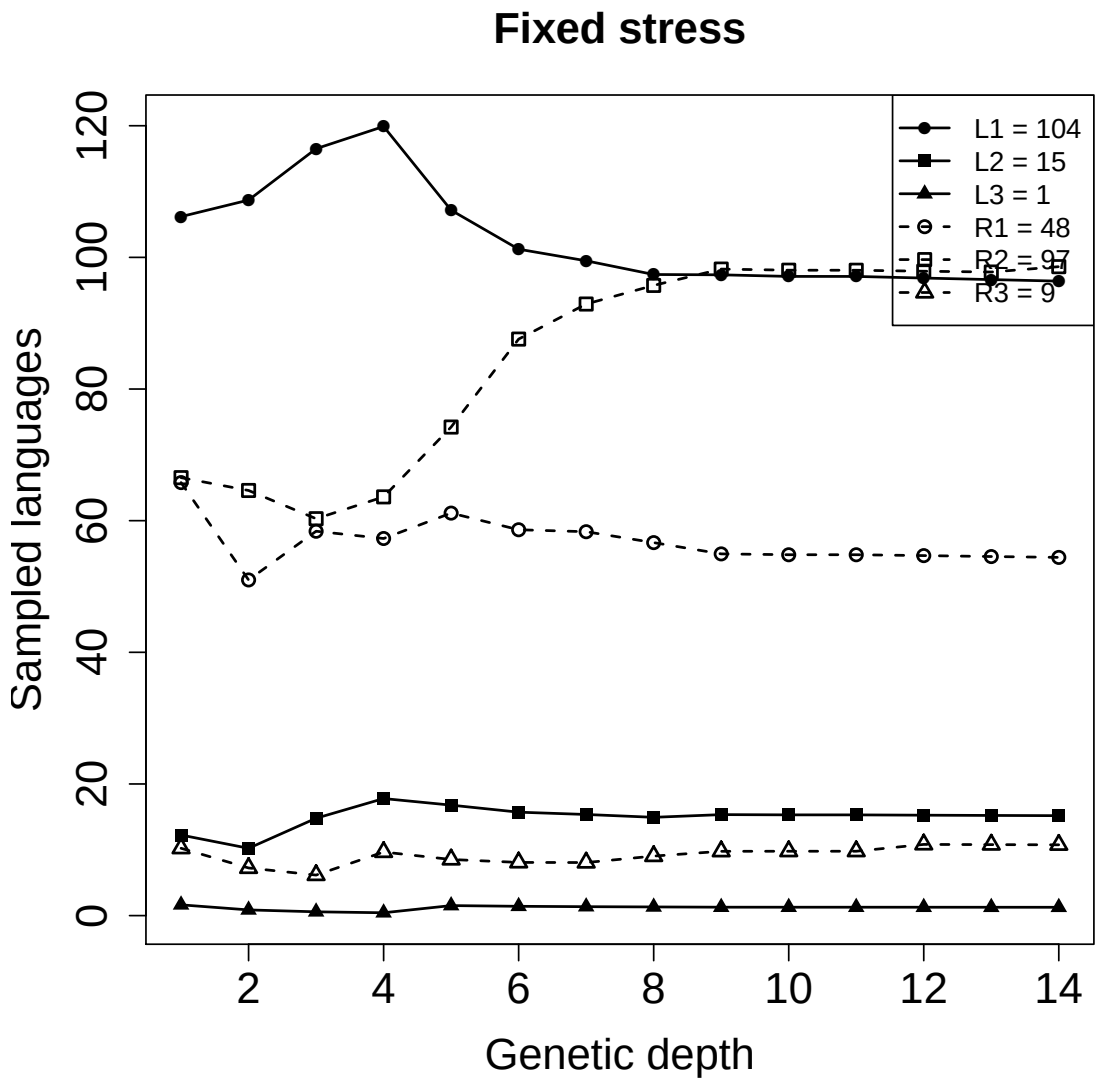


Figure A.8: Number of fixed stress languages found in 50,000 resamples of the data, uniformly sampled within a genetic depth. L1, L2, L3 are initial, peninitial, and postpeninitial stress. R1, R2, and R3 are final, penultimate, and antepenultimate stress. Legend shows observed values.

Some genetic biases would not be analyzable with these methods. This would be the case if a large number of genetic classifications were completely unsampled in StressTyp *and* those missing classifications exhibited different overall patterns than seen in StressTyp. This is not possible to know for sure, but for a typology of this size it is not unreasonable to suppose that the missing classifications are missing at random, not systematically.

Finally, I make no study here of areal biases. These are undoubtedly also present, as the frequency of Australian languages attests. The process for biased sampling of area groupings is not immediately obvious, but uniform sampling could be achieved by sampling random points on a sphere and comparing these to summary latitude and longitude data, as present in StressTyp2 (van der Hulst, 2014). I leave this analysis for future work.

BIBLIOGRAPHY

Adam, Galit, and Outi Bat-El. 2009. When do universal preferences emerge in language development? The acquisition of Hebrew stress. *Brill's Annual of Afroasiatic Languages and Linguistics* 1:255–282.

Alber, Birgit. 2005. Clash, lapse and directionality. *Natural Language & Linguistic Theory* 23:485–542.

Alderete, John. 1995. Winnebago accent and Dorsey's Law. In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*, ed. Jill Beckman, Laura Walsh-Dickey, and Suzanne Urbanczyk, 21–51. Amherst, MA: Graduate Linguistic Student Association.

Andersson, Jonas, and Jan Ubøe. 2012. Some aspects of random utility, extreme value theory and multinomial logit models. *Stochastics: An International Journal of Probability and Stochastic Processes* 84:425–435.

Asher, Ronald E., and T.C. Kumari. 1997. *Malayalam*. Psychology Press.

Bach, Emmon, and Robert T. Harms. 1972. How do languages get crazy rules. *Linguistic change and generative theory* 1:21.

Bailey, Todd M. 1995. Nonmetrical constraints on stress. Doctoral Dissertation, University of Minnesota.

Bane, Max, and Jason Riggle. 2008. Three correlates of the typological frequency of quantity-insensitive stress systems. In *Proceedings of Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, volume 10, 29–38. Association for Computational Linguistics.

Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22:39–71.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, 43–58.

Boersma, Paul. 2003. Review of Tesar & Smolensky (2000): Learnability in Optimality Theory. *Phonology* 20:436–446.

Boersma, Paul, and Joe Pater. 2014. Convergence properties of a gradual learner in Harmonic Grammar. In *Harmonic Grammar and Harmonic Serialism*, ed. John J. McCarthy and Joe Pater. London: Equinox Press.

Chomsky, Noam. 1979. *Principles and parameters in syntactic theory*.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*.

Coetzee, Andries W. 2002. Between-language frequency effects in phonological theory.

Crowhurst, Megan J., and Mark S. Hewitt. 1995. Directional footing, degeneracy, and alignment. In *Proceedings of the North East Linguistics Society*, ed. Jill Beckman, volume 25, 47–61. Graduate Linguistic Student Association.

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

Dell, François, and Mohamed Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics* 7:105–130.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society* 39:1–38.

Dixon, Robert M.W. 1983a. Nyawaygi. In *Handbook of Australian languages*, ed. Robert M.W. Dixon and Barry J. Blake, 431–531. Amsterdam: John Benjamins.

Dixon, Robert M.W. 1983b. Wargamay. In *Handbook of Australian languages*, ed. Robert M.W. Dixon and Barry J. Blake. Amsterdam: John Benjamins.

Everett, Dan, and Keren Everett. 1984. On the relevance of syllable onsets to stress placement. *Linguistic Inquiry* 705–711.

Facundes, Sidney da Silva. 2000. The language of the Apurinã people of Brazil (Maipure/Arawak). Doctoral Dissertation, State University of New York.

Goedemans, Rob. 2010. A typology of stress patterns. In *A survey of word accentual systems in the language of the world*, ed. Harry van der Hulst, Rob Goedemans, and Ellen van Zanten, 647–666. Berlin: Mouton de Gruyter.

Goedemans, Rob, and Harry van der Hulst. 2013. Fixed stress locations. In *The world atlas of language structures online*, ed. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL `http://wals.info/chapter/14`.

Goedemans, Rob, Harry van der Hulst, and Ellis Visch. 1996a. Part 1: Background. In *Stress patterns of the world*, ed. Rob Goedemans, Harry van der Hulst, and Ellis Visch. The Hague: Holland Academic Graphics.

Goedemans, Rob, Harry van der Hulst, and Ellis Visch. 1996b. StressTyp: A database for prosodic systems in the worlds languages. *Glot International* 2:21–23.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, 111–120.

Golston, Chris. 1996. Direct Optimality Theory: Representation as pure markedness. *Language* 72:713–748.

Gordon, Matthew. 2000. The tonal basis of weight criteria in final position. In *Regional Meetings, Chicago Linguistic Society*, volume 36, 141–156.

Gordon, Matthew. 2002. A factorial typology of quantity-insensitive stress. *Natural Language & Linguistic Theory* 20:491–552.

Griffiths, Thomas L., and Michael L. Kalish. 2005. A Bayesian view of language evolution by iterated learning. In *Proceedings of the annual conference of the cognitive science society*, volume 27, 827–832.

Griffiths, Thomas L., and Michael L. Kalish. 2007. Language evolution by iterated learning with Bayesian agents. *Cognitive science* 31:441–480.

Hammond, Michael. 1991. Parameters of metrical theory and learnability. *Logical issues in language acquisition* 47–62.

Harris, Alice C, and Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*, volume 74. Cambridge University Press.

Hayes, Bruce. 1985. Iambic and trochaic rhythm in stress rules. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, ed. Vassiliki Nikiforidou Mary Niepokuj, Mary VanClay and Deborah Feder, volume 11.

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hayes, Bruce P. 1999. Phonetically driven phonology. *Functionalism and formalism in linguistics* 1:243–285.

Heinz, Jeffrey. 2007. Inductive learning of phonotactic patterns. Doctoral Dissertation, University of California, Los Angeles.

Hintz, Diane M. 2006. Stress in South Conchucos Quechua: a phonetic and phonological study. *International journal of American linguistics* 72:477–521.

van der Hulst, Harry. 1996. Separating primary and secondary accent. In *Stress patterns of the world*, ed. Rob Goedemans, Harry van der Hulst, and Ellis Visch. The Hague: Holland Academic Graphics.

van der Hulst, Harry. 2014. *Word stress: Theoretical and typological issues*. Cambridge University Press.

Hyde, Brett. 2008. The rhythmic foundations of initial gridmark and nonfinality. In *Proceedings of the North East Linguistics Society*, volume 38.

Hyman, Larry. 1977. On the nature of linguistic stress. *Studies in stress and accent* 4.

Hyman, Larry M. 1985. *A theory of phonological weight*. Foris Publications Dordrecht.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. *Architectures, rules, and preferences: a festschrift for Joan Bresnan* 467–479.

Jarosz, Gaja. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology* 30:27–71.

Jensen, John Thayer, Leo David Pugram, John Baptist Iou, and Raphael Defeg. 1977. *Yapese reference grammar*. University Press of Hawaii Honolulu.

Jusczyk, Peter W, Anne Cutler, and Nancy J Redanz. 1993. Infants' preference for the predominant stress patterns of English words. *Child development* 64:675–687.

Kager, René. 2001. Rhythmic directionality by positional licensing.

Kager, René. 2005. Rhythmic licensing theory: an extended typology. In *Proceedings of the third international conference on phonology*, 5–31. Seoul National University.

Kager, René. 2012. Stress in windows: Language typology and factorial typology. *Lingua* .

Karvonen, Daniel. 2008. Explaining nonfinality: Evidence from Finnish. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 26, 306–314.

Kennedy, Chris. 1994. Morphological alignment and head projection. *Phonology at Santa Cruz* 3:47–64.

Kenstowicz, Michael. 1993. Peak prominence stress systems and Optimality Theory. In *Proceedings of the First International Conference on Linguistics at Chosun University*, 7–22.

Kirby, Simon. 2002. Learning, bottlenecks and the evolution of recursive syntax. *Linguistic evolution through language acquisition: Formal and computational models* 173–203.

Kirby, Simon, Mike Dowman, and Thomas L. Griffiths. 2007. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences* 104:5241–5245.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. *Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness*. University of Colorado, Boulder, Department of Computer Science.

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006. The Optimality Theory-Harmonic Grammar connection. In *The harmonic mind: From neural computation to Optimality Theoretic grammar, volume 2: Linguistic and philosophical implications*, ed. Paul Smolensky and Géraldine Legendre, 339–402. Cambridge, MA: MIT Press.

Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic inquiry* 8:249–336.

Littlestone, Nick. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning* 2:285–318.

Lunden, Stephanie Laura. 2006. Weight, final lengthening and stress: A phonetic and phonological case study of norwegian. Doctoral Dissertation.

MacWhinney, Brian. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Maddieson, Ian. 1980. Phonological generalizations from the UCLA Phonological Segment Inventory Database (UPSID). In *UCLA working papers in linguistics*, volume 50, 57–68.

McCarthy, John J. 1979. On stress and syllabification. *Linguistic inquiry* 10:443–465.

McCarthy, John J. 2003. Ot constraints are categorical. *Phonology* 20:75–138.

McCarthy, John J., and Alan Prince. 1993a. Generalized alignment. *Yearbook of morphology* 79–153.

McCarthy, John J., and Alan Prince. 1993b. Prosodic morphology I: Constraint interaction and satisfaction. Technical report, Rutgers University Center for Cognitive Science.

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25:83.

Moreton, Elliott, and Joe Pater. 2012. Structure and substance in artificial-phonology learning. *Language and linguistics compass* 6:686–701.

Moreton, Elliott, Joe Pater, and Katya Pertsova. in prep. Phonological concept learning.

Niyogi, Partha, and Robert C Berwick. 2009. The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences* 106:10124–10129.

Novikoff, Albert B.J. 1962. On convergence proofs for perceptrons. In *Proceedings of the symposium on the mathematical theory of automata*, volume 12.

Oller, D. Kimbrough. 2005. The effect of position in utterance on speech segment duration in English. *The journal of the Acoustical Society of America* 54:1235–1247.

Padgett, Jaye. 1995. Feature classes. In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*, 385–420.

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

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

Pater, Joe. 2014. Universal grammar with weighted constraints. In *Harmonic Grammar and Harmonic Serialism*, ed. John J. McCarthy and Joe Pater. London: Equinox Press.

Pater, Joe, and Elliott Moreton. 2012. Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad* 3:1–41.

Pater, Joe, and Robert Staubs. 2013. Feature economy and iterated grammar learning. *Manchester Phonology Meeting 22. Presentation.* .

Payne, Judith. 1990. Asheninca stress patterns. *Amazonian linguistics, ed. by Doris Payne* 185–212.

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

Prince, Alan. 1993. *In defense of the number i: Anatomy of a linear dynamical model of linguistic generalizations*. Rutgers Center for Cognitive Science.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Malden, MA and Oxford, UK: Blackwell.

Prince, Alan S. 1983. Relating to the grid. *Linguistic inquiry* 15:19–100.

Pruitt, Kathryn Ringler. 2012. Stress in Harmonic Serialism. Doctoral Dissertation, University of Massachusetts Amherst.

Rafferty, Anna N., Thomas L. Griffiths, and Marc Ettlinger. 2011. Exploring the relationship between learnability and linguistic universals. *Association for Computational Linguistics: Human Language Technologies 2011* .

Rafferty, Anna N, Thomas L Griffiths, and Dan Klein. 2009. Convergence bounds for language evolution by iterated learning. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Riggle, Jason. 2008. Counting rankings.

Rosenblatt, Frank. 1957. *The perceptron, a perceiving and recognizing automaton project para*. Cornell Aeronautical Laboratory.

Smalley, William A. 1953. Phonemic rhythm in Comanche. *International journal of American Linguistics* 19:297–301.

Smith, Jennifer. 2004. *Phonological augmentation in prominent positions*. Routledge.

Smolensky, Paul, and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic grammar*. MIT Press.

Staubs, Robert. 2011. Operational Exponence: Process morphology in Harmonic Serialism. *Challenges of Complex Morphology to Morphological Theory, Linguistic Society of America Summer Institute, Boulder, CO. Handout* .

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. The MIT Press.

Theisen, Carrie Ann, Jon Oberlander, and Simon Kirby. 2010. Systematicity and arbitrariness in novel communication systems. *Interaction studies* 11:14–32.

Vihman, Marilyn May, Rory A. DePaolis, and Barbara L. Davis. 1998. Is there a trochaic bias in early word learning? Evidence from infant production in English and French. *Child development* 69:935–949.

Wedel, Andrew. 2011. Self-organization in phonology. *The Blackwell Companion to Phonology. Blackwell* 130–147.

Weide, Robert L. 1994. CMU pronouncing dictionary. URL `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Widrow, Bernard, and Marcian E. Hoff. 1960. Adaptive switching circuits. Technical report, Defense Technical Information Center.

Zuraw, Kie. 2003. Probability in language change. *Probabilistic linguistics* 139–176.