Learning bias in stress windows: Frequency and attestation

Robert Staubs
University of Massachusetts Amherst

1 Introduction

In a traditional generative framework, relative frequency is not directly modeled. Some patterns are treated as "possible," others as "impossible." No gradation is encoded as an explicit part of the model. This focus is problematic: relative frequency exists, and therefore should be explained in some way. However, if a grammar with standard assumptions is combined with an account of learning, it is able to model frequencies. In this paper, I provide such a combined account, modeling both relative frequency and gaps in the typology of stress windows. This result is derived by integrating an explicit consideration of learning into the process of typological modeling.

In a stress window system, stress is required to fall within a given number of syllables from an edge. Within this "window," the choice of stress position is dictated by some designated property of the syllable or word. This property might be quantity (weight), sonority, or lexically marked stress. I will refer to this property as a designated property, without theoretical commitment to its representational interpretation. An example of stress window data is given in Table 1. Kobon shows a sonority-driven stress within a final two-syllable window. Within that window, the highest-sonority vowel bears stress. If a high sonority vowel falls farther away from the word edge, a syllable within the window is stressed instead. Stress windows show several marked frequency asymmetries—two of principal interest for this paper. First, windows are more common when they are small. Second, windows of size four or larger are not attested. Counts for these systems are shown in Table 2.

In this paper, I show that the relationship between window size and frequency is not surprising given learning considerations. First, the size of a window increases, larger and larger words are needed to detect it. This is problematic, given that long words are comparatively rare. Second, as noted by Prince (1993:p. 12), the learning problem in a weighted grammatical model makes larger windows more difficult to acquire. As window length increases, a narrower and narrower range of weights describes a pattern. Analogously, the 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, much of the 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. I show that this bias emerges from an iterated learning model using an online learner of Maximum Entropy grammar, incorporating current approaches to phonological grammar and learning to explicitly model frequency.

This type of bias from learning is important for two reasons. Most crucially, it opens up a concrete approach to a type of data that is often worryingly unaddressed, namely relative frequency. In addition, a frequency model offers an approach to questions of attestation. In categorical generative frameworks, a grammatical theory that permits an unattested pattern is undesirable. Such an argument forms the basis of an objection by Legendre et al. (2006) to weighted grammars for typology: they (righly) 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. Long windows are not readily generated in typical OT constraint sets (Kager, 2012)—but neither are small windows. This type

* This material is based upon work supported by the National Science Foundation under Grant No. S121000000211. Thanks to audiences at the following related talks: LSA 2015, AMP 2014, mfm 22, RUMMIT IV, UMass Phonology Reading Group (Fall 2013), Sound Workshop (Spring 2014), dissertation defense (Summer 2014). Thanks also to Gillian Gallagher, René Kager, John Kingston, Nazarré Merchant, John McCarthy, and Joe Pater for various insights.

Proceedings of *AMP 2014*
Completed February 26, 2015

ki.ˈa	‘tree species’	[a] beats [i] on sonority
ˈhau.i	‘vine species’	[a] beats [i] on sonority
ga.ˈkʷi.nə	‘bird species’	[a] too far from edge, [i] beats [ə]
a.ˈla.go	‘snake species’	stress the [a] in the window

Table 1: Kobon shows a sonority-driven final two-syllable window (Davies, 1981; Kenstowicz, 1994).

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

Table 2: 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.

of theoretical objection, based on gaps, is considerably weakened in a frequency view of typology: if a given pattern (here, long stress windows) is predicted to be low frequency, it should not be surprising if this is manifested as a typological gap. Thus, models with “undesirable” categorical predictions can be attractive when used to predict relative frequency.

In this paper, I first give the type of model used. Next, I give representative simulation results and discuss in more detail how typological bias emerges from such a model. Finally, I show these results are connected to relative frequency and attestation of stress systems through iterated learning, and conclude.

2 Model

2.1 *Grammar* In the model used in this paper, grammars are represented as weighted constraints, implemented as a Maximum Entropy grammar (MaxEnt; Goldwater & Johnson, 2003). The constraints of MaxEnt are similar to those of Optimality Theory (Prince & Smolensky, 1993/2004): each constraint assigns a candidate some number of violations due to the structure of the candidate. MaxEnt differs first in that the constraints are weighted (assigned real number values), not ranked. This means that MaxEnt is a type of Harmonic Grammar (HG; Legendre et al., 2006; Pater, 2009), and is therefore subject to cumulativity effects. MaxEnt is distinguished from some other HG theories in that candidates are assigned probabilities—there is not just a single winner and a set of losers. MaxEnt assigns probabilities in proportion to the exponential of harmony—the weighted sum of constraint violations.

The constraints used in this study were chosen to represent standard sorts of distinctions in stress grammars, modifying the constraint sets of Alber (2005) and Kager (2005) specifically and McCarthy & Prince (1993) and Prince & Smolensky (1993/2004) more generally.

1. ALIGNHEADLEFT/RIGHT Violated for every syllable between the main stress and left/right edge.
2. **“**WEIGHTTOSTRESS**”** (WSP) Violated for every designated property-bearing syllable without stress.
3. FTBIN Violated for every degenerate foot.
4. *LAPSE Violated for every sequence of unstressed syllables.
5. IAMB/TROCHEE Violated for every left/right-headed foot.

In truth, only WSP and ALIGN are necessary for the effects discussed in this paper. However, the remaining constraints were included in these simulations as part of a larger set of stress learning simulations (Staubs, 2014a).

$\sigma\sigma\underline{\sigma}\sigma$	WSP 15	ALIGN-R 10	H	p
$\sigma\sigma\acute{\sigma}\sigma$		-1	-10	0.99
$\sigma\sigma\sigma\acute{\sigma}$	-1		-15	0.01

Table 3: Final two-syllable window: evaluation of a designated property that falls within the window. An underline marks the position of the designated property in the word.

$\sigma\underline{\sigma}\sigma\sigma$	WSP 15	ALIGN-R 10	H	p
$\sigma\acute{\sigma}\sigma\sigma$		-2	-20	0.01
$\sigma\sigma\sigma\acute{\sigma}$	-1		-15	0.99

Table 4: Final two-syllable window: evaluation of a designated property that falls outside the window. An underline marks the position of the designated property in the word.

We can use these two constraints and the cumulativity of MaxEnt to model basic stress windows (Legendre et al., 2006). Table 3 gives a weighting of the two crucial constraints yielding a final two-syllable window. The weights of 10 and 15, multiplied by the violation scores of -1, give harmonies of -10 and -15 to the two candidates. When these harmonies are exponentiated and normalized, the resulting probabilities are 0.99 and 0.01. Thus, when the designated property falls within the window, the effect of WSP dominates and the syllable (almost always) bears stress. Crucially, these probabilities are internal to a language; they do not establish the typological probabilities of particular patterns.

When the designated property falls outside the window, as in Table 4, the violations of ALIGN "gang up" on WSP. There are now two violations of ALIGN but still only a single WSP violation. The order of harmonies and probabilities thus flips: in this situation, a default (final) position is selected for stress.

The objection raised by Legendre et al. (2006) is derived from this sort of analysis. If a two-syllable window can be represented this way, with a ratio of 15 to 10, a three-syllable window can be represented by a ratio of 25 to 10 or a four-syllable window with a ratio of 35 to 10. This process could be extended indefinitely to represent windows as large as desired, in defiance of the typological observation that windows are at most three syllables long. One purpose of this paper is to demonstrate that this type of prediction is not necessarily a problem when learning is considered. To do this, we need an explicit incorporation of learning into the model.

2.2 *Learning* The learning model used here is an online one—the learner processes each datum as it is received from the teacher, updating its grammar as it goes. The process has essentially four parts:

1. **Sampling:** Teacher produces a form according to its (categorical) grammar.
2. **Interpretation:** Learner attempts to reconstruct the hidden structure for the teacher's presentation.
3. **Production:** Learner (probabilistically) produces its parse.
4. **Update:** If there is a mismatch, the learner updates its constraint weights.

2.2.1 *Sampling* Each teacher has a fixed grammar representing a stress pattern of interest, for example a final two-syllable window. In sampling, the teacher chooses a form to give to the learner. In this model, this is done randomly. A teacher first decides which word length to produce. To maintain a connection between the learning data and actual real-world data, this choice obeys an exponentially decaying function. To be precise, the probability of selecting a given word length n is proportional to 2^{-n}. The particular

choice of exponential decay does not appear to be important—in fact, I show that results are similar with the (unrealistic) assumption of uniform sampling. The teacher selects lengths between two and eight syllables.

After the selection of a word length, a single designated property is potentially assigned to the word. This is done uniformly at random: one syllable is chosen to bear the property, with no bias for or against any position in the word. Word shapes without a designated property are also possible—this is counted as a single syllabic position. Thus the probability of a designated property on any given syllable is $(n+1)^{-1}$, as is the probability of no designated property.

2.2.2 *Interpretation* As each teacher has a particular fixed grammar, at this point the teacher can parse the word and give its resulting stress pattern to the learner. The learner receives only the stresses and the position of a designated property—not the hidden foot structure. This is not enough for the learner to go on: it must know the violations of the teacher's form with respect to foot constraints. The learner therefore needs some way to guess at the teacher's intentions. The approach presented here uses a probabilistic adaptation of Robust Interpretive Parsing (RIP; Tesar & Smolensky, 2000; Boersma, 2003; Jarosz, 2013; Boersma & 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. That is, compatible structures are chosen with probability proportional to the exponential of their harmony. 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.

2.2.3 *Production* In production, the learner generates its own parse of the teacher's given word shape. This is done without respect to the teacher's chosen stress pattern: consideration is given only to the length of the word and the position of the designated property (if any). Given the probabilistic nature of MaxEnt grammars, this production will not always be the same. Each parse is just a sample from the learner's grammatical understanding of the given word shape. For present purposes, the learner considers only candidates that have a single stress.

2.2.4 *Update* The learner updates its grammar by comparing its own output to the form received from the teacher and its guess about the teacher's hidden structure. The learner infers the constraint violations of the teacher's form using Robust Interpretive Parsing. These are compared with the violations of the learner's form. The weights are updated in proportion to the difference between the two violation vectors. Any constraint violated more by the teacher than the learner will go down in weight, while any constraint violated less by the teacher will go up in weight. This serves to move probability from the learner's current output onto its interpretation of the teacher's form, if these are different. This method is the MaxEnt sampling Stochastic Gradient Ascent (SGA, perceptron, HG-GLA; Jäger, 2007; Boersma & Pater, 2014).

(1) $\text{New Weights} = \text{Old Weights} + \eta\,(\text{Learner Violations} - \text{Teacher Violations})$

Where η is some assumed learning rate, assumed throughout to be equal to 0.1. Initial weights are drawn from a normal distribution ($\mu = \sigma = 10$) truncated to ≥ 0. The minimum weight is 0, and update beyond this results in a zero weight instead.

3 Learning bias

3.1 *Simulation results* The performance of a learner can be computed by comparing the learner's distribution over parses to the distribution of the teacher. A simple way to make this comparison is sum squared error (SSE). With sum squared error, the probability of a form under the teacher's (categorical) grammar is compared to the learner's (probabilistic) grammar. These are subtracted, squared, and summed up. A learner that disagrees with its teacher—that has not learned much—will have a high SSE, one that has learned the teacher's grammar will have a zero SSE.

SSE^t is the SSE of the learner after t words given by the learner. Two learners can be compared by comparing SSE^t for the same t. If the SSE is lower for one learner, it has learned more in the same amount of time. The relative ease of learning for Language 1 versus Language 2 can therefore be computed by averaging the SSE for many learners with a teacher of Language 1 with a similar average for learners with a teacher of Language 2.

1.	$\underline{\sigma}\sigma$	$\rightarrow$	$\acute{\sigma}\sigma$
2.	$\underline{\sigma}\sigma\sigma$	$\rightarrow$	$\acute{\sigma}\sigma\sigma$
3.	$\underline{\sigma}\sigma\sigma\sigma$	$\rightarrow$	$\acute{\sigma}\sigma\sigma\sigma$
4.	$\underline{\sigma}\sigma\sigma\sigma\sigma$	$\rightarrow$	$\sigma\sigma\sigma\sigma\acute{\sigma}$

Table 5: Ambiguity of a long window: final four-syllable window not clear in words under five syllables.

Simulations were performed using the model from §2. Briefly restated, this model has an online, error-driven learner using a MaxEnt grammar of metrical stress constraints. The learner attempts to learn a categorical stress window language. Long words are sampled exponentially more than short words (2^{-n}) and designated properties are distributed uniformly. Words of length two to eight are considered, with the candidates being all corresponding parses with one stress. The learning rate η was held at 0.1 for all simulations.

Figure 1a shows the average SSE of learners under these conditions. As the number of iterations (datapoints given to the learner) increases, SSE decreases. That is, the learners learn the target distribution for the three patterns considered: two-, three-, and four-syllable windows with an edge default. At every point in learning, however, smaller windows have a lower average SSE than larger ones. This means that small windows are better learned with fixed data or, equivalently, learned equally well with less data.

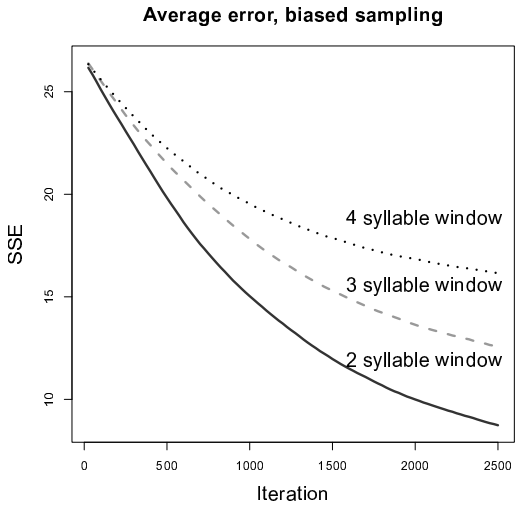
These results show that the model is in fact biased, but the source of this bias is ambiguous. It could come from the size of words required for learning, from the reliability of constraint violations, or both. Figure 1b shows results for a model in which the teacher samples words of length two to eight uniformly. This removes the real-world bias towards short words, but leaves a distinction between the three stress systems. Figure 1c goes further, showing results for a simulation in which learners were only provided words of length eight. In this case it is no longer possible for word length *per se* to affect the results at all. Instead, only reliability and the distribution of designated properties are relevant. It must be noted, however, that these graphs show a faster rate of learning overall than the exponentially-biased case—word length is not irrelevant, it is simply not the whole picture.

3.2 *Explaining bias* These simulation results show that the model presented in §2 is biased toward short windows. Why should this be? There are two important reasons. First, short windows do not need the learner to encounter long words. Second, short windows are (relatively) consistent across word lengths.

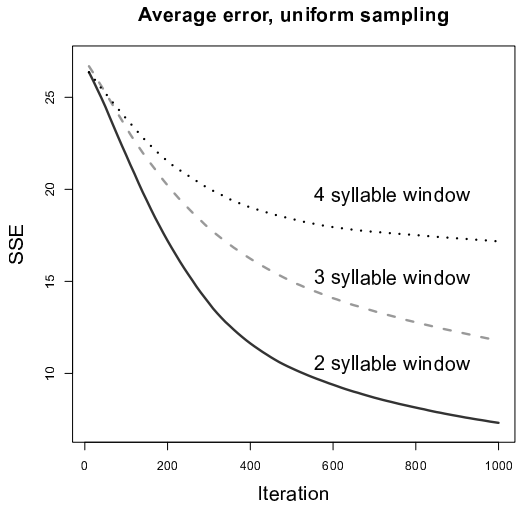
Very long stress windows require very long words in order to disambiguate them from other conceivable stress systems. In short words, a long window placed at the right edge looks the same as a long left-edge window, unbounded stress, etc. This is reflected in Table 5. This table shows a final four-syllable window. In words under five syllables long, stress falls wherever the designated property is. It is not clear that this is a four-syllable window; it is only clear that stress tends to fall on the designated property. It is only when five-syllable forms are revealed that the learner—or, indeed, the analyst—could discover the "true" system, even in principle. Given that these forms are rare, we expect such systems to be less learnable. Similar considerations are advanced elsewhere, for example by Bailey (1995) and Bane & Riggle (2008).

The previous explanation is based on the structure of the input data. An additional explanation is available based on the structure of the learner itself. A given learning model will be biased in the way it learns—some things will be learned quickly, others more slowly (if at all). The learning model presented here is biased towards reliable constraint violations. The task of the learner is to separate out the tolerable violations found in the forms preferred by the teacher from the unacceptable violations found in all other logically possible forms. This is represented clearly in the update rule: weights change in a way proportional to the difference between the learner's violations and the teacher's.

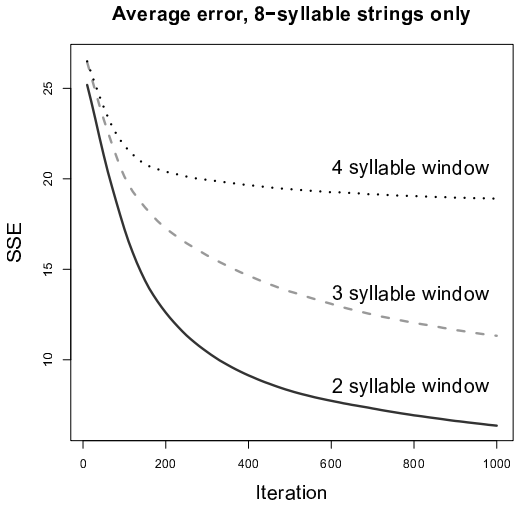
Learning is fastest when updates tend to point in the same direction. It is difficult to separate the teacher's forms from others if there is a lot of variation in the teacher's violations: every bit of variation is a case in which the target forms look more like the "noise" of all other forms and less like a cohesive whole. If the teacher's forms did not vary at all, updates would progressively march in the direction of the teacher's



(a) Exponentially decreasing word length distribution.



(b) Uniform length distribution. Note axis change.



(c) Only words of length eight.

Figure 1: SSE for learners of window languages.

grammar as quickly as possible. This idea relates to considerations for the perceptron algorithm discussed by Novikoff (1962).

Short windows offer this situation of reliability, while long windows become increasingly unreliable. Table 6 gives the five-syllable strings of languages of varying window sizes. Each system is assumed to be right-aligned with a final default. Stress tracks with the designated property as it moves away from the edge until the size of the window is exceeded. Table 7 shows the types of ALIGN violation tolerated in each of these languages. In a two-syllable window, ALIGN can be violated either zero times or one time. A three-syllable window allows zero violations, one, or two. A four-syllable window can tolerate three violations, and so on. Each increase in the size of the window adds inconsistency to the position of stress and decreases the reliability of constraint violations. As the size of windows increase, therefore, learning will be harder and harder as the updates become more and more spread out.

	σσσσσ	σσσσσ	σσσσσ	σσσσσ	σσσσσ
2-syllable	σσσσ́	σσσσ́	σσσσ́	σσσσ́	σσσσ́
3-syllable	σσσσ́	σσσ́σ	σσ́σσ	σσσσ́	σσσσ́
4-syllable	σσσσ́	σσσ́σ	σσ́σσ	σ́σσσ	σσσσ́
5-syllable	σ́σσσσ	σσσ́σ	σσ́σσ	σ́σσσ	σσσσ́

Table 6: Basic window stress patterns.

	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) Two-syllable.

	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) Three-syllable.

	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) Four-syllable.

	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**	**4**			
6	0	**1**	**2**	**3**	**4**	0		
7	0	**1**	**2**	**3**	**4**	0	0	
8	0	**1**	**2**	**3**	**4**	0	0	0

(d) Five-syllable.

Table 7: Patterns of violations of ALIGN across window sizes for same-edge default final windows. 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.

As seen in §3.1, the distribution over input forms and the distribution over violations within those forms both play a role in biasing learning toward short windows. If either is removed, a bias still remains, but it is diminished. With this relative difference in learning for long and short stress windows, we can move towards a model of the relative frequency of such systems.

4 Iterated learning and typology

To address typological predictions, it is not enough to examine only the relative rate of language learning as done above. This process is informative, but could be misleading about how typology might be shaped by learners with this kind of bias (see discussion by Rafferty et al., 2011). A more direct approach incorporates a schematic model of historical change through (mis)learning. In an iterated learning model (e.g. Kirby, 2002; Griffiths & Kalish, 2005) learners serve as teachers for the next generation of learners (schematized in Figure 2). Over many generations, the biases in learning influence the likelihood of particular hypotheses (i.e. language patterns).

For the purposes of this work, iterated learning is viewed through the probabilities of a single generation of language transmission. Again, learners attempt to learn from categorical teachers of various stress patterns. After some amount of learning (2,500 iterations—words given to the learner by the teacher[1]), learning is terminated. At this point, the resulting distribution of the learner is compared to all relevant teacher distributions. The closest distribution (by maximum likelihood) is determined to be the language learned by the learner. When this process is carried out many times (to determine statistics) for every teacher language under consideration, a distribution over learning outcomes is produced. This result describes how likely a learner is to arrive at any given language hypothesis for any given teacher language. If learners always quickly learned the languages of their teachers, this would not be an interesting procedure, but instances of "failed" learning open the way to iterated learning calculations.

The languages of interest chosen here 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, but it allows easy comparison between language. 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 3 shows bias in learning outcomes. Fixed stress 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.

This is the beginning of a learning bias in typology—large windows are not typically learned as faithful versions of themselves. However, it is still not enough. This is still not a model of the dynamics present in a full iterated learning system. To do this, we must take a single generation of learning, as represented here, and project it forward across many generations. When the statistics are explicitly represented, as here, this is a simple process: estimates for future generations are derived by exponentiating the previous result. Figure 4 shows the 100th power of Figure 3—the predicted statistics after 100 generations. The near-categorical outcomes of fixed stress do not change. Stress windows, however, consolidate probability toward lower syllable counts. Two-syllable windows are maintained, while larger windows are learned predominantly as three- or four-syllable windows.

This consolidation effect increases across successive generations. Figure 5 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. 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. The early rise of three-syllable windows is due to their

[1] "Iteration" is used in this paper only to refer to an individual interaction between a teacher and a learner—the transmission of a single piece of data and its corresponding update. "Generation" is used to one full set of iterations in transmitting language from one teacher to one learner.

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

Figure 2: Schematic view of iterated learning.

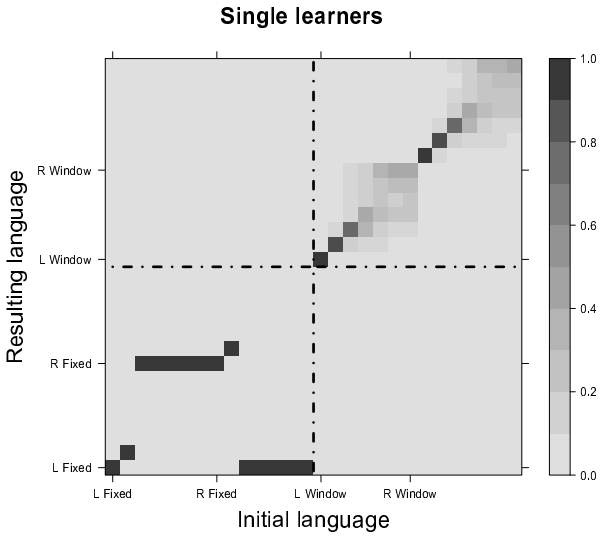


Figure 3: Single step learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language. Within each labeled category, number of syllables in the window increases to the right. $\eta = 0.1$, 2,500 trials per language, 2,500 words given by teacher to learner for each.

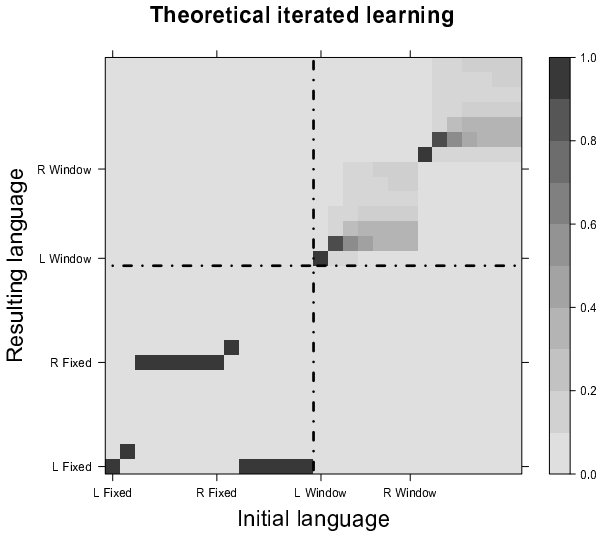


Figure 4: Simulated iterated learning confusion matrix. Probability of ending at some language after starting at some (possibly different) language estimated over 100 trials. Within each labeled category, number of syllables in the window increases to the right. $\eta = 0.1$, 2,500 words given by teacher to learner for each.

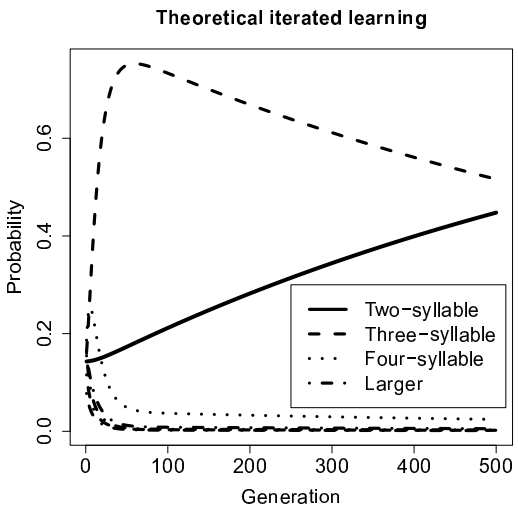


Figure 5: Dominance of small windows in predicted iterated learning. Proportion of stress windows taken up by a certain size across predicted generations. Calculated from learning results of Figure 3. $\eta = 0.1$.

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.[2]

The model does well at distinguishing the relative frequency of different window sizes. The question of attestation and maximal window size is less clear. These models are indeed capable of predicting some number of four-syllable windows. However, the predicted probability of such a system is not necessarily high. Figure 6 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? 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 observed zero count is by no means unlikely (20.60% probability).[3] This model is therefore plausible in that it could easily produce an "accidental" gap (no four-syllable windows) that has been characterized elsewhere as a principled gap.

It is worth stressing that the model does not predict that these exact numeric values must hold of the real world. Instead, I have chosen the generation at which the model gives its best approximation to the real-world data. In one sense, this is a strange choice. Such an iterated learning model forms a Markov chain with a stationary distribution (e.g. Griffiths & Kalish, 2007). As an inherent characteristic of the chain, this would be a natural choice for evaluation. However, such a focus would make two unwarranted assumptions about the world. First, that language transmission has "converged" to a stationary distribution, which is perhaps unlikely (Rafferty et al., 2009). Second, that the relative learning of just this set of languages is the *only* factor

[2] 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.

[3] 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.

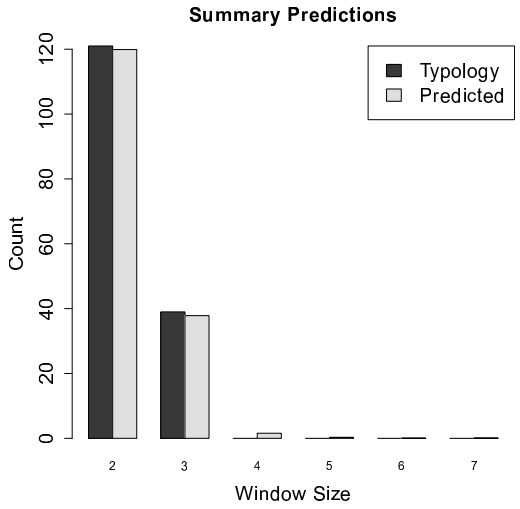


Figure 6: Comparison of predicted frequencies with typology (numbers from Kager, 2012). Counts sum over left and right windows. Exponent 1,664 chosen by minimizing sum squared error with data.

in language change, excluding contact, reanalysis, other sources of "weight-driven" stress, etc. In the face of these issues, I ask only whether the model is *compatible* with real-world observations, finding that it appears to be.

5 Conclusion

In this paper, I have presented a learning-based model of frequency in the typology of stress windows. Such models predict relative levels of attestation for stress languages on the basis of the reliability of constraint violations and the mutual reinforcement of stress data. In particular, through comparison of residual learning error and iterated learning, I show that short windows are preferred by a plausible set of stress constraints situated within an error-driven learner of MaxEnt grammar. This preference for short windows follows the preference observed in the typology of window stress. In addition, I demonstrated that this type of model is compatible with the observed absence of windows of size four and above.

These simulations provide validation for a learning-based approach to explaining a large part of the tendencies in window stress typology. The frequency results are not obtainable at all 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. This type of approach is promising for stress typology in general: Staubs (2014b,a) applies these same methods to other aspects of stress, while Stanton (2014) uses a related approach to provide a response to the midpoint pathology that does not need modified constraints.

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—combined with a learning theory automatically yields 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. Ultimately, breaking down the wall separating grammatical formalisms from their learning yields a richer and more natural theory of language.

References

Alber, Birgit (2005). Clash, lapse and directionality. *Natural Language & Linguistic Theory* 23:3, 485–542.

Asher, Ronald E. & T.C. Kumari (1997). *Malayalam*. Psychology Press.

Bailey, Todd M. (1995). *Nonmetrical constraints on stress*. Ph.D. thesis, University of Minnesota.

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

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

Boersma, Paul & Joe Pater (2014). Convergence properties of a gradual learner in Harmonic Grammar. McCarthy, John J. & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, Equinox Press, London.

Davies, John (1981). Kobon. Lingua descriptive series, vol. 3.

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

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

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

Griffiths, Thomas L. & Michael L. Kalish (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive science* 31:3, 441–480.

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 & Raphael Defeg (1977). *Yapese reference grammar*. University Press of Hawaii Honolulu.

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

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

Kenstowicz, Michael (1994). Sonority-driven stress. *ROA-33* .

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

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

McCarthy, John J. & Alan Prince (1993). Generalized alignment. *Yearbook of morphology* 79–153.

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

Pater, Joe (2009). Weighted constraints in generative linguistics. *Cognitive Science* 33:6, 999–1035.

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

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

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

Smalley, William A. (1953). Phonemic rhythm in Comanche. *International journal of American Linguistics* 19:4, 297–301.

Stanton, Juliet (2014). Learnability shapes typology: the case of the midpoint pathology. *lingbuzz/002347* .

Staubs, Robert (2014b). Learning and the position of primary stress. *Proceedings of the 31st West Coast Conference on Formal Linguistics*, 428–437.

Staubs, Robert D. (2014a). *Computational modeling of learning biases in stress typology*. Ph.D. thesis, University of Massachusetts Amherst.

Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. The MIT Press.

Theisen, Carrie Ann, Jon Oberlander & Simon Kirby (2010). Systematicity and arbitrariness in novel communication systems. *Interaction studies* 11:1, 14–32.