Learning Stress with Feet and Grids*

Seung Suk Lee, Cerys Hughes, Alessa Farinella, and Joe Pater
University of Massachusetts Amherst

1 Introduction

Foot-based approaches to stress introduce representational ambiguity in that a single stress pattern can correspond to multiple prosodifications (e.g. batáma as (batá)ma or ba(táma), where parentheses show foot boundaries). This case of structural ambiguity, or hidden structure, has been a major focus of research on phonological learning, including the foundational work in Dresher & Kaye (1990) and Tesar & Smolensky (2000). Tesar and Smolensky (henceforth TS) propose an approach to hidden structure learning in which the learner's current grammar is used to prosodify learning data that consist of strings of light and heavy syllables with stress indicated, without any prosodic structure. They test their approach on a test set of 124 'languages' that have a variety of quantity sensitive and quantity insensitive stress patterns. Subsequent research, including Boersma & Pater (2016) and Jarosz (2013, 2015) has explored the consequences of changes to the grammatical and learning assumptions of TS for performance on that test set.

This paper follows up on Pater & Prickett (2022) (henceforth PP), who find that many of the stress patterns that their MaxEnt learner tends to fail on in the TS test set are in fact unattested ('MaxEnt' abbreviates Maximum Entropy, a probabilistic weighted constraint grammar model; see Sections 2 and 3). To begin to assess how well the grammar and learning model adopted in that work, as well as others, can cope with languages that are claimed to be typologically attested, we studied the learning of a test set of 33 languages based on the Gordon (2002) typology of quantity insensitive stress, as presented in Hayes & Wilson (2008). Our foot-based constraint set, which is similar but not identical to the one in TS, can represent 26 of the 33 languages in the Gordon typology. We found that a grammar+learning model based on the one in PP was able to find a correct grammar for all of these languages, and that none would be classified as 'hard' according to the PP measure of difficulty.

We also compared the performance of the learner equipped with a foot-based constraint set to the same learner with two grid-based constraint sets based on Gordon (2002). These constraint sets assume representations that are in a one-to-one relationship with strings of syllables with stress designations. Therefore, there is no structural ambiguity with respect to the learning problems we are studying here, and the learning algorithm we are employing is guaranteed to find a correct grammar for every language that the constraints can represent. Instead of success rate, our comparisons are in terms of the number of epochs (learning steps) that it takes to find a correct grammar. We find that the foot-based learner is faster in these terms than the learner equipped with the unmodified constraint set from Gordon (2002) (though as we discuss, it is slower in real time). The second grid-based learner changes the main stress constraint to count syllables rather than secondary stresses between the main stress and the word edge, as in the version of the main stress constraint used for our foot-based constraints. The mean number of epochs to convergence for this revised grid-based learner is about the same as the foot-based learner.

We conclude that the foot-based learner and the grid-based learner fare similarly well in this initial comparison on a typologically grounded set of learning problems. Much further research, both in terms of developing constraint sets and in testing them in learning, is needed to assess how well these types of learners fare on a broader typology, and with more realistic assumptions about the learning data.

* We thank Brandon Prickett for providing the MaxEnt learner and for his valuable feedback, all the people at the UMass Sound Workshop, especially Gaja Jarosz, six anonymous AMP reviewers, and participants in AMP 2022. Thanks also to Bruce Hayes for providing the training data for the learning simulations. This research was supported by NSF grant BCS-2140826 to UMass Amherst.

Proceedings of *AMP 2022*

2 Constraint sets and learning problems

Like the constraint set in TS, our foot-based constraint set draws on the proposals in Prince & Smolensky (1993/2004) and McCarthy & Prince (1993). It differs from TS in a few ways. First, it adopts the standard formulation of the constraint preferring initially stressed feet, which we call 'TROCHAIC'. As PP discuss, the TS constraint that prefers trochees, which also penalizes monosyllabic feet, does not allow for an analysis of Latin and similar quantity sensitive languages. Second, our 'NONFINALITY' constraint assigns an extra penalty for final stress, following Prince & Smolensky (1993/2004). And third, we adopt a proposal from Kager (2005) to eliminate the constraints that Tesar and Smolensky call 'WORDFOOTR' and 'WORDFOOTL', which penalize words whose final or initial syllables respectively are not footed.

(1) Foot-based constraint set

FTBIN: Assign a violation for each foot that consists only of a light syllable.

PARSE: Assign a violation for each syllable that is not in a foot.

IAMBIC: Assign a violation for each initially stressed disyllabic foot.

TROCHAIC: Assign a violation for each finally stressed disyllabic foot.

NONFINALITY: Assign one violation if the final syllable of a word is footed, and a second violation if the final syllable is also stressed.

MAINR: Assign a violation for every syllable intervening between the syllable with main stress and the right edge of the word.

MAINL: Assign a violation for every syllable intervening between the syllable with main stress and the left edge of the word.

ALLFEETR: Assign a violation for every syllable intervening between the right edge of each foot with the right edge of the word.

ALLFEETL: Assign a violation for every syllable intervening between the left edge of each foot with the left edge of the word.

These constraints can generate most, but not all of the patterns in Gordon's typology of quantity insensitive stress. One notable example of a pattern that they can capture is one in which primary stress falls on the final syllable, and secondary stress falls on the initial, even in disyllabic words, instantiated as Québec French in the typology. Gordon (2011) cites this pattern as problematic for a foot-based system. It may not be at first obvious how this constraint set can account for it, since if it is an iambic system as the final stress would seem to indicate, why should the initial syllable of a disyllable be stressed as in the correct $(\grave{\sigma})(\acute{\sigma})$ rather than parsed as the weak member of an iambic foot, leading to the incorrect $*(\sigma\acute{\sigma})$? An analysis is available with trochaic feet, since the correct form can be generated with MAINR preferring final main stress in $(\grave{\sigma})(\acute{\sigma})$ over a single trochaic foot as in the incorrect $*(\acute{\sigma}\sigma)$.

One language in Gordon's typology that cannot be generated by this constraint set is the Indonesian stress pattern as provided in Hayes & Wilson (2008); the description is attributed to Cohn (1989) in Gordon (2002). This pattern, like Garawa which we will discuss in detail in Section 5, involves what is termed an initial dactyl. As we discuss in that section Indonesian escapes analysis in our constraint set because of the elimination of WORDFOOTL. Four other languages that our constraint set fails to capture are the ones that involve ternary iteration: Cayuvava, Ioway-Oto, Pacific Yupik, and Winnebago. These are also not generated by the Tesar and Smolensky constraints, and the expansions of the constraint set needed to capture them are non-trivial and controversial (see recently Martínez-Paricio & Kager (2015)). Finally, two of the languages in the Hayes & Wilson (2008) instantiation of the Gordon typology have multiple stress patterns for a given string of syllables. As we are following TS in studying only deterministic stress placements we also omit those two, Estonian and Walmatjari. This leaves 26 languages in our test set.

We have two grid-based constraint sets. The first is adopted directly from Gordon (2002). The constraints are as follows, with definitions rephrased so as to correspond to ones we give for the foot-based constraints. In the grid-based representation assumed here, syllables are mapped to level 0 grid marks. Level 1 grid marks indicate secondary stress, and level 2 grid marks indicate primary stress.

(2) Grid-based constraint set (Gordon (2002))

ALIGN(X_1, L, 0, PrWd) (ALIGN1LPRWD): Assign a violation for every grid mark at level 0 intervening between each grid mark at level 1 and the left edge of the word.

ALIGN(X_1, R, 0, PrWd) (ALIGN1RPRWD): Assign a violation for every grid mark at level 0 intervening between each grid mark at level 1 and the right edge of the word.

ALIGN(EDGES, level 0, PrWd, X_1) (ALIGNEDGES): Assign a violation for every grid mark on the edges at level 0 that does not have a grid mark at level 1 (maximum 2 violations per word).

NONFINALITY: Assign one violation if the final syllable has a grid mark at level 1.

*CLASH: Assign a violation for every pair of consecutive syllables that have a level 1 grid mark.

*LAPSE: Assign a violation for every pair of consecutive syllables lacking a level 1 grid mark.

*EXTENDEDLAPSE: Assign a violation for every triplet of consecutive syllables lacking a level 1 grid mark.

*LAPSERIGHT: Assign one violation mark if there is more than one level 0 grid mark (≥ 2) intervening between the rightmost level 1 grid mark and the right edge of the word.

*LAPSELEFT: Assign one violation mark if there is more than one level 0 grid mark (≥ 2) intervening between the leftmost level 1 grid mark and the left edge of the word.

*EXTENDEDLAPSERIGHT: Assign one violation mark if there are more than two level 0 grid marks (≥ 3) intervening between the rightmost level 1 grid mark and the right edge of the word.

ALIGN(X_2, L, 1, PrWd) (ALIGN2LPRWD): Assign a violation for every grid mark at level 1 intervening between each grid mark at level 2 and the left edge of the word.

ALIGN(X_2, R, 1, PrWd) (ALIGN2RPRWD): Assign a violation for every grid mark at level 1 intervening between each grid mark at level 2 and the right edge of the word.

Our second grid-based constraint set is the same as (2) except that the two ALIGN(X_2, L/R, 1, PrWd) constraints are revised as in (3).

(3) ALIGN(X_2, L, 0, PrWd) (ALIGN2LPRWD$_\sigma$): Assign a violation for every grid mark at level 0 intervening between each grid mark at level 2 and the left edge of the word. (i.e. just count the syllables)

ALIGN(X_2, R, 0, PrWd) (ALIGN2RPRWD$_\sigma$): Assign a violation for every grid mark at level 0 intervening between each grid mark at level 2 and the right edge of the word. (i.e. just count the syllables)

These constraints do not follow the general formulation for ALIGN constraints in Gordon (2002), which requires constraints on level 2 (main stress) to refer only to level 1 (secondary stress). Though formulated in terms of grids rather than feet, they assign violation marks in the same way as the MAINL/R constraints in the foot-based constraint set, by counting syllables rather than secondary stresses between the main stress and the word edge.

The learning problems we investigate are similar in structure to those set up by TS. The learner is provided with a constraint set and must find a ranking (or weighting) that generates the correct form for each tableau. When there is hidden structure, correctness is defined in terms of what TS call the 'Overt' form. For the TS stress problems, and the ones we investigate here, the Overt form is a string of syllables with stress designations. In Table 1, we illustrate how the learning problem is set up for both a grid- and a foot-based learner. The 'LD' column (for Learning Data) indicates which of the Overt forms is correct; the example here specifies that the correct form has final main stress and no secondary stress. In the foot-based table on the right, there are two prosodifications that correspond to the correct Overt form, shown in the 'Full Rep' (Full Representation) column. For learning to be successful in TS, one of these full representations must be made optimal. In our probabilistic approach, we follow PP in requiring that the sum of the probabilities of the correct full representations to be greater than 0.9. For conciseness, we omit the grid-based full representations, since they map one-to-one with the Overt representations. The tables also

GRID					FOOT					
Input	Overt	LD	*CLASH	...	Input	Overt	LD	Full Rep	FTBIN	...
σσ	σσ́	1	0	...	σσ	σσ́	1	(σσ́)	0	...
								σ(σ́)	1	...
	σ́σ	0	0	...		σ́σ	0	(σ́σ)	0	...
								(σ́)σ	1	...
	σ́σ̀	0	1	...		σ́σ̀	0	(σ́)(σ̀)	2	...
	σ̀σ́	0	1	...		σ̀σ́	0	(σ̀)(σ́)	2	...

Table 1: Left: Example of a two-syllable word for learning with grid-based constraints. Right: Example of a two-syllable word for learning with foot-based constraints

provide an example of a constraint assigning a violation count to each candidate: *CLASH for the grid-based constraint set and FTBIN for the foot-based one.

The learning data for each language is a set of strings from 1 to 7 syllables in length, marked for degree of stress (primary, secondary, or none). These are the data used in Hayes and Wilson's (2008) MaxEnt phonotactic learning simulations for the Gordon (2002) typology, provided by Hayes (p.c.). Our approach differs from Hayes and Wilson's in using predefined, typologically grounded constraints, in allowing for hidden structure, and mapping from inputs to outputs. For each of the seven-syllable lengths, there is a tableau with competing candidate stress patterns, like those in Table 1. The Overt forms include every possible stress pattern where one and only one syllable bears primary stress. For the foot-based candidate sets, every compatible prosodification with maximally bisyllabic feet is provided for each Overt form.

3 Learning algorithm and test procedure

Like PP, we adopt Maximum Entropy Grammar, a probabilistic version of Prince and Smolensky's Optimality Theory (1993/2004) proposed by Goldwater & Johnson (2003). Constraints are given numerical weights rather than ranks, and the probability of a candidate is proportional to the exponentiated weighted sum of constraint violations. We also adopt PP's approach to hidden structure, which uses a version of Expectation Maximization first applied by Pater et al. (2012) to other hidden structure problems.

Our learner differs from the PP one in that we use Gradient Descent to update constraint weights, rather than the L-BFGS-B optimization algorithm. In pilot work, L-BFGS-B was found to outperform Gradient Descent in learning the TS languages (Brandon Prickett, p.c.). Gradient Descent has the advantage of providing smoother, more interpretable learning trajectories. We use the batch version of Gradient Descent described in Pater & Staubs (2013) and Moreton et al. (2017), in which each update is over the entire dataset.

We used the same learning procedure for the three constraint sets. For each one of the 26 languages, we did a run with the constraint weights initialized at 1, and 10 runs with each constraint weight sampled with replacement from a uniform distribution of 0-10. We track learning in terms of the number of updates, or epochs, that it takes the learner to reach our criterion for having a correct grammar, which as we noted above is that the correct stress pattern in every tableau must be given at least 0.90 probability. In cases of hidden structure, we sum over the prosodifications that yield the correct stress pattern. For example, in the foot-based learning problem in Table 1, we check if the probabilities that are assigned to '(σσ́)' and 'σ(σ́)', which are both compatible with the winner 'σσ́', sum up to at least 0.90. Weights were kept positive by replacing any negative weight with zero (see Magri (2015)). The learning rate was set to 4 (a value found to work well in pilot work on the TS languages), and there was a maximum of 1000 epochs.

4 Overall results

When the weights were initialized at 1 the learner equipped with the foot-based constraint set (henceforth FOOT) found a correct set of weights for all 26 languages. For the 10 random initializations, the learner found a correct grammar except in a minority of runs for the three languages in (4). The number of failed runs is shown in parentheses. We discuss the failed runs for Garawa in the next section.

(4) Languages with failed runs for the FOOT learner
Garawa (3), Georgian (2), Southern Paiute (1)

PP classified TS languages as 'hard' if their learner failed with initialization at 1, or if it failed on all 10 random initializations. By that standard, none of the languages in the current study is hard. This success rate is particularly remarkable since we used Gradient Descent rather than L-BFGS-B.

As expected, our learner equipped with Gordon's grid-based constraints (henceforth GRID) succeeded on every run. There was similar uniform success for the version of the grid-based learner whose main stress constraint counts syllables rather than secondary stresses between the main stress and the word edge (henceforth GRID-MAIN-REV). This shows that the reformulation of that constraint does not change the constraint set's ability to capture the typology (it also succeeds on the Indonesian and the ternary languages that were left out of our test set).

	GRID	GRID-MAIN-REV	FOOT
Mean	56.0	30.2	34.5
Median	18.5	11.5	12.5
Min	1	1	1
Max	402	260	143

Table 2: Epochs to criterion with initial weights = 1

Table 2 provides summary statistics for the number of epochs required to reach the correctness criterion with initialization at 1. The learner using Gordon's constraint set (GRID in Table 2) takes on average many more epochs to find a correct analysis than does the foot-based learner. This cannot be taken as a general finding about learning with feet and grids however, as shown by the rather dramatic drop in number of epochs to criterion for GRID-MAIN-REV. This learner's mean and median number of epochs to criterion are in fact slightly lower than FOOT, though its maximum is nearly twice as high. Table 3 shows that this general pattern holds up with random initialization (FOOT runs include only the successful ones). The Min and Max rows show the mean number of epochs to criterion for a specified language over ten runs. The difference on the maximum number between GRID-MAIN-REV and FOOT is even greater here, since the FOOT maximum drops to 94 from 143.

	GRID	GRID-MAIN-REV	FOOT
Mean	47.6	29.9	29.8
Median	22.2	12.6	17.8
Min	1.0	1.0	1.0
Max	360.7	258.5	94.0

Table 3: Mean of 10 runs with randomly initialized weights

There are likely two reasons that GRID-MAIN-REV is so much faster than GRID. First, there is the simple fact that a count of intervening syllables will usually be greater, and will never be smaller, than a count of intervening secondary stresses, which will result in a generally larger update of the GRID-MAIN-REV constraint. Second, as Gaja Jarosz (p.c.) points out, the extent to which the weight of GRID main stress constraint is changed depends on how probable candidates with intervening secondary stress are; this does not affect the GRID-MAIN-REV constraint.

In sum, we found that the foot-based learner and the grid-based learner generally took a similar number of epochs to find a correct grammar when the main stress constraint was made to apply in the same way in the two systems. It terms of real time though, the grid-based learner is considerably quicker. When we ran the learner implemented in Python (https://github.com/blprickett/Hidden-Structure-MaxEnt.git) on Google Colab GRID type learning took approximately 45 seconds to go through 1000 epochs while FOOT learning took 187 seconds, which is 4 times longer than GRID. This is due to the greater complexity of calculations for the foot-based system brought about by the greater number of candidates. The sizes of the candidate sets are compared in Table 4.

Number of Syllables	Grid	Foot	Grid/Foot
1	1	1	1
2	4	6	0.67
3	12	24	0.50
4	32	88	0.36
5	80	300	0.27
6	192	984	0.20
7	448	3136	0.14
All	769	4539	0.17

Table 4: Number of candidates for grid and foot-based systems

Ultimately, rather than overall efficiency, the most useful comparison amongst the learners may well focus on the relative speed at which different languages are learned by each one. The graphs in Figure 1 provide an initial comparison that shows that different constraint sets often have different relative speeds of learning the languages. For each constraint set, the number of epochs to criterion of each language (with initialization at one) was log transformed and normalized. The graphs show the three pairwise comparisons amongst our three constraint sets. Each dot represents a language, and dots close to the line are ones whose relative difficulty is similar across the pair of constraint sets. We have labeled some of the languages. Atayal and Chitimatcha have final and initial main stress respectively, with no secondary stress, and both are learned in 1 epoch by all three constraint sets. The comparison of the grid-based constraint sets in the rightmost graph has the dots closest to the line, indicating that the relative difficulty of the languages is closest for them. Lakota has peninitial main stress, and no secondary stress, and both grid-based learners find it relatively difficult. That relative difficulty is much higher than for the foot-based learner, as shown in the position of Lakota in the leftmost and center graphs. Malak Malak, which has stress on even-numbered syllables counting from the right, and primary stress on the leftmost of those, is an example of a language with the reverse distinction, with relative difficulty being higher for the foot-based learner. Arriving at a better understanding of the biases of the learners, and comparing them to typological frequencies or human learning data is a clear potential direction for future research. Some initial comparison of grid- and foot-based learners along these lines can be found in Staubs (2014b) and Staubs (2014a). The difficulty that a foot-based learner has with the type of language illustrated by Malak Malak is the focus of Staubs' work, and it is interesting in that context that Malak Malak was relatively easy for the revised grid constraint set.

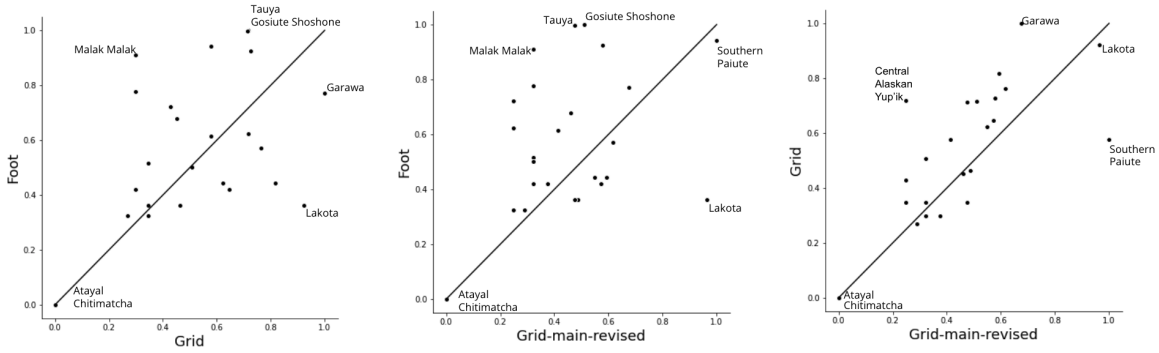

Figure 1: Number of epochs when weights are initialized at 1, log transformed and normalized within each constraint set. Axis labels indicate the constraint set used in learning, and each dot represents a language

5 Results on Garawa

The language that FOOT failed on most often was Garawa, in 3 out of 10 random initializations. Interestingly, this language displays a type of ambiguity that is somewhat similar to the unattested and difficult to learn languages that are the focus of the PP study that we build on here. In Table 5, the stress patterns that are the target of learning are shown in the 'Observed' column. An analysis that uses only trochaic feet is shown in 'Trochaic Analysis'. The 'Mixed Analysis' uses a combination of iambic and trochaic feet. The words of 2 and 6 syllables of length are shown with two prosodifications. The probabilities shown in 'Mixed Analysis' are the ones that are assigned to these outcomes by the grammar learned with one of the random initializations. In the Mixed Analysis, the final syllable is unparsed because Nonfinality has a relatively high weight. It is this ambiguity between a final stressless syllable being unparsed or being parsed as the dependent of a trochaic foot that leads to failures of learning in the unattested languages discussed by PP, and as we will see, to the failed runs for Garawa here.

Number of σ	Observed	Trochaic Analysis	Mixed Analysis	Mixed Probabilities
1	$\acute{\sigma}$	$(\acute{\sigma})$	$(\acute{\sigma})$	1
2	$\acute{\sigma}\sigma$	$(\acute{\sigma}\sigma)$	$(\acute{\sigma}\sigma)$	0.53
			$(\acute{\sigma})\sigma$	0.47
3	$\acute{\sigma}\sigma\sigma$	$(\acute{\sigma}\sigma)\sigma$	$(\acute{\sigma}\sigma)\sigma$	0.99
4	$\acute{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\acute{\sigma})(\sigma\grave{\sigma})\sigma$	0.94
5	$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})\sigma$	0.97
6	$\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\acute{\sigma})(\sigma\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.89
			$(\acute{\sigma}\sigma)(\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.10
7	$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.93

Table 5: Garawa learning data and analyses

Table 6 shows the constraint weights for the trochaic analysis that was found by the learner with constraint weights initialized at 1. As mentioned above, Garawa is an example of what is called an initial dactyl language: in strings of 5 and 7 syllables of length, there is a two-syllable lapse separating the initial stress from the next one. The fixed stress on the initial syllable is due to the relatively high weight of MAINL. In particular, the correct $(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$ is preferred over $*\sigma(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$ because the weight of MAINL is greater than that of ALLFEETRIGHT. This ability of the syllable-counting MAINL/R constraints to pin the main stress at a fixed distance from an edge is the reason that Kager (2005) was able to eliminate constraints from McCarthy & Prince (1993) that align a single foot to the edge of the word, termed WORDFOOT by TS. Most of the time, in languages with alternating stress oriented to one edge, and a fixed stress at the other, the fixed stress is the main stress. Indonesian, which has an initial dactyl and main stress on the rightmost of the stresses is an exception. It is worth noting that the empirical facts for Indonesian stress are unclear; the initial dactyl pattern described by Cohn (1989) has not been instrumentally verified – see Goedemans & van Zanten (2007) and Athanasopoulou et al. (2021).

Observed	Parses	Probabilities	Constraints	Learned weights	Initial weights
$\acute{\sigma}$	$(\acute{\sigma})$	1.00	PARSE	10.26	1
$\acute{\sigma}\sigma$	$(\acute{\sigma}\sigma)$	1.00	MAINL	8.17	1
$\acute{\sigma}\sigma\sigma$	$(\acute{\sigma}\sigma)\sigma$	1.00	FTBIN	6.91	1
$\acute{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	1.00	ALLFEETRIGHT	5.90	1
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)$	0.99	TROCHAIC	4.08	1
$\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	1.00	NONFINALITY	2.02	1
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)\sigma(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	1.00	ALLFEETLEFT	1.39	1
			MAINR	0	1
			IAMBIC	0	1

Table 6: Weights for trochaic analysis found with initialization at 1; reached criterion in 46 epochs

Table 7 shows the initial and final weights for the mixed analysis from table 5. As can be seen in the 'Initial weights' column, NONFINALITY started out with the highest weight of the constraints; IAMBIC was

lower, but above TROCHAIC. This learner got to the correct analysis very quickly: after only 3 epochs. The two parses for the disyllabic word have about equal probability because the sums of the weights of the constraints that each violate are about equal. The NONFINALITY and IAMBIC violations in $(\acute{\sigma}\sigma)$ trade off against the PARSE and FTBIN violations in $(\acute{\sigma})\sigma$.

Observed	Parses	Probabilities	Constraints	Learned weights	Initial weights
$\acute{\sigma}$	$(\acute{\sigma})$	1.00	MAINL	13.92	5.32
$\acute{\sigma}\sigma$	$(\acute{\sigma}\sigma)$	0.53	NONFINALITY	9.68	9.53
	$(\acute{\sigma})\sigma$	0.47	PARSE	9.27	8.00
$\acute{\sigma}\sigma\sigma$	$(\acute{\sigma}\sigma)\sigma$	0.99	FTBIN	6.46	7.81
$\acute{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma})(\sigma\grave{\sigma})\sigma$	0.94	IAMBIC	5.92	6.68
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})\sigma$	0.97	TROCHAIC	2.58	1.78
$\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma})(\sigma\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.89	ALLFEETLEFT	1.07	1.91
	$(\acute{\sigma}\sigma)(\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.10	MAINR	0.13	8.73
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.93	ALLFEETRIGHT	0	1.04

Table 7: Weights for mixed analysis found with random initialization; reached criterion in 3 epochs

Table 8 shows a failed run, typical of all three failed runs. Here again NONFINALITY is helping to keep stress off the final syllable, allowing iambs to be used as the non-initial feet in the highest probability parses for the 5 and 7 syllable words, just as they are in the correct analysis. Those words do not reach the criterion of 0.90 correct because too much probability is being given to non-initial trochaic feet; the weight of IAMBIC is not sufficiently greater than that of TROCHAIC. In the 6 syllable word, the problem is reversed: the medial iamb places stress in the wrong position. The trap seems to be caused by high-weighted FTBIN: the correct analysis in Table 7 uses monosyllabic feet in the six-syllable word. The failed runs all had initial weights with IAMBIC higher than TROCHAIC, and differed from the successful run in Table 7 in having a relatively low NONFINALITY weight.

Observed	Parses	Probabilities	Constraints	Learned weights	Initial weights
$\acute{\sigma}$	$(\acute{\sigma})$	1.00	PARSE	12.76	8.77
$\acute{\sigma}\sigma$	$(\acute{\sigma}\sigma)$	1.00	FTBIN	10.14	9.36
$\acute{\sigma}\sigma\sigma$	$(\acute{\sigma}\sigma)\sigma$	1.00	MAINL	9.14	1.85
$\acute{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)$	1.00	NONFINALITY	8.28	0.22
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})\sigma$	0.74	IAMBIC	4.87	9.60
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	0.25	TROCHAIC	3.78	1.44
$\acute{\sigma}\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)$	0.25	ALLFEETRIGHT	1.36	2.19
	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})(\grave{\sigma}\sigma)$	0.75	ALLFEETLEFT	0.87	6.84
$\acute{\sigma}\sigma\sigma\grave{\sigma}\sigma\grave{\sigma}\sigma$	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})(\sigma\grave{\sigma})\sigma$	0.56	MAINR	0	1.96
	$(\acute{\sigma}\sigma)(\sigma\grave{\sigma})(\grave{\sigma}\sigma)\sigma$	0.19			
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\sigma\grave{\sigma})\sigma$	0.19			
	$(\acute{\sigma}\sigma)(\grave{\sigma}\sigma)(\grave{\sigma}\sigma)\sigma$	0.06			

Table 8: Weights for a failed run with random initialization, candidates that are not compatible with the observed data are in gray

6 Conclusions

Foot-based MaxEnt learning was remarkably successful in this initial investigation of the learning of a typologically based test set of languages. With initialization of the constraints at 1, a correct grammar was found for all of the 26 languages of the Gordon (2002) typology that the constraint set can represent, and the learner rarely failed to find a correct grammar when weights were randomly initialized. Foot-based learning was also as efficient as grid-based learning, at least when speed was measured in terms of the number of epochs to criterion. This leads us to tentatively conclude that the avoidance of hidden structure in the grid-based approach does not provide a learning argument for it, *pace* Gordon (2011).

Firmer conclusions await an expansion of the typology being covered by the constraint sets. As discussed above, the foot-based constraint set that we adopted cannot cover the entirety of the Gordon typology, and neither of these constraint sets is designed to cover the full range of typologically attested patterns, including also quantity sensitive ones. And in studying quantity sensitive patterns, there is a further hidden structure problem of identifying which of the syllables are 'heavy' and 'light', which TS and others abstract from. In building up to a fuller typology, there are also other learning options to explore, including Expectation Driven Learning as presented in Jarosz (2015)) and Nazarov & Jarosz (2021), and theories that do not assume a prespecified constraint set, such as the MaxEnt constraint induction model of Hayes & Wilson (2008), and the neural network approach of Prickett & Pater (2022). Finally, going beyond success and overall speed of learning, there is a need for further study of the biases of each set of grammar and learning assumptions, in terms of the relative ease of different language types, and how the learning space is navigated on the way to success.

References

Athanasopoulou, Angeliki, Irene Vogel & Nadya Pincus (2021). Prosodic prominence in a stressless language: An acoustic investigation of Indonesian. *Journal of Linguistics* 57:4, 695–735.

Boersma, Paul & Joe Pater (2016). Convergence properties of a gradual learner in Harmonic Grammar. McCarthy, John J. & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, Equinox Publishing, Bristol, Connecticut, 389–434.

Cohn, Abigail C (1989). Stress in Indonesian and bracketing paradoxes. *Natural Language & Linguistic Theory* 7:2, 167–216.

Dresher, B. Elan & Jonathan D. Kaye (1990). A computational learning model for metrical phonology. *Cognition* 34:2, 137–195, URL `https://linkinghub.elsevier.com/retrieve/pii/001002779090042I`.

Goedemans, Rob & Ellen van Zanten (2007). Stress and accent in Indonesian. *LOT Occasional series* 9, 35–62.

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

Gordon, Matthew (2002). A factorial typology of quantity-insensitive stress. *Natural Language & Linguistic Theory* 20:3, 491–552.

Gordon, Matthew (2011). Stress systems. *The Handbook of Phonological Theory* 75.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:3, 379–440.

Jarosz, Gaja (2013). Learning with hidden structure in optimality theory and harmonic grammar: Beyond robust interpretive parsing. *Phonology* 30:1, 27–71.

Jarosz, Gaja (2015). Expectation driven learning of phonology. *Ms. University of Massachusetts Amherst* .

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

Magri, Giorgio (2015). How to keep the HG weights non-negative: the truncated Perceptron reweighing rule. *Journal of Language Modelling* 3:2, 345–375.

Martínez-Paricio, Violeta & René Kager (2015). The binary-to-ternary rhythmic continuum in stress typology: layered feet and non-intervention constraints. *Phonology* 32:3, 459–504, URL `http://www.jstor.org/stable/43865864`.

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

Moreton, Elliott, Joe Pater & Katya Pertsova (2017). Phonological concept learning. *Cognitive Science* 41:1, 4–69, URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12319`.

Nazarov, Aleksei Ioulevitch & Gaja Jarosz (2021). The credit problem in parametric stress: A probabilistic approach. *Glossa: a journal of general linguistics* 6:1, URL `https://doi.org/10.16995/glossa.5884`.

Pater, Joe & Brandon Prickett (2022). Typological gaps in iambic nonfinality correlate with learning difficulty. *Proceedings of the Annual Meetings on Phonology*, vol. 9.

Pater, Joe & Robert Staubs (2013). Modeling learning trajectories with batch gradient descent. *NECPhon, MIT* .

Pater, Joe, Robert Staubs, Karen Jesney & Brian Cantwell Smith (2012). Learning probabilities over underlying representations. *Proceedings of the twelfth meeting of the Special Interest Group on Computational Morphology and Phonology*, 62–71.

Prickett, Brandon & Joe Pater (2022). Learning Stress Patterns with a Sequence-to-Sequence Neural Network. *Proceedings of the Society for Computation in Linguistics* 5:10, URL `https://scholarworks.umass.edu/scil/vol5/iss1/10/`. Publisher: University of Massachusetts Amherst.

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

Staubs, Robert (2014a). *Computational modeling of learning biases in stress typology*. Ph.D. thesis, University of Massachusetts Amherst, URL http://scholarworks.umass.edu/dissertations$_2$/230/.

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

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