Typological Gaps in Iambic Nonfinality Correlate with Learning Difficulty

Joe Pater and Brandon Prickett
University of Massachusetts Amherst

1 Overview

In this paper, we discuss gaps in stress typology that are unexpected from the perspective of a foot-based theory and show that the patterns pose difficulties for a computationally implemented learning algorithm. The unattested patterns result from combining theoretical elements whose effects are generally well-attested, including iambic footing, nonfinality, word edge alignment and a foot binarity requirement. One of these patterns is particularly descriptively simple: stress is final in disyllables that begin with a light syllable (and in monosyllables), and penultimate otherwise. The fact that it, and a set of related patterns, pose a great difficulty to a class of learning algorithms suggests that their cross-linguistic absence may be due to learning factors.

These patterns can be found amongst the 124 target stress systems constructed by Tesar and Smolensky (2000; henceforth TS) as a test of their approach to hidden structure learning. They do not flag the systems' typological unattestedness, nor does any of the subsequent literature that uses this test set (Boersma 2003; Jarosz 2013; 2015; Boersma and Pater 2016). We became aware of their cross-linguistic absence when we found that they were part of a set of systems that our own learner failed on and looked to the theoretical-typological literature to see if they existed.

Our learner operates with a Maximum Entropy grammar (MaxEnt; Goldwater and Johnson 2003), and uses a form of Expectation Maximization to deal with hidden structure, as in Pater et al. (2012) (see Jarosz 2006; 2013; 2015 for other applications of Expectation Maximization in phonology). This is the first study of Maximum Entropy learning of this well-studied test set; we find that the success rate is just slightly lower than the state-of-the-art results presented in Jarosz (2013) and Jarosz (2015).

2 Two simple but unattested stress systems

We begin by discussing two of the unattested stress systems from the TS test set that present difficulties to our learner (and related ones), so as to motivate this study from the perspective of stress typology. The first is the one already briefly mentioned: stress the final syllable of a bisyllabic word with a light initial syllable, else stress the penultimate syllable. This pattern falls out straightforwardly from standard constraints in Optimality Theory (OT; Prince and Smolensky 2004). The table in (1) shows a comparison of candidates that the analysis makes optimal, or Winners, with relevant competing candidates, or Losers (Prince 2002). The *L* in a candidate refers to a light syllable, an *H* to a heavy one, and a following *1* indicates that the syllable is stressed. Parentheses show foot boundaries. The three constraints are standard OT stress constraints (McCarthy and Prince 1993; Prince and Smolensky 2004) that are used by TS in their test set (see sec. 4 below for definitions). FtBin is violated by a foot consisting only of a light syllable, Nonfin is violated by a word-final syllable that is footed, and AllFtR assigns a violation for every syllable separating each foot from the right edge of the word. A *W* beneath a constraint indicates that it prefers the Winner in that row, while an

* We thank David Smith and Robert Staubs for discussion, as well as members of the UMass Sound Workshop, especially Gaja Jarosz, and participants in AMP 2021, especially Bruce Hayes. Thanks also to Paul Boersma for sharing the Tesar and Smolensky (2000) test set, and files needed to replicate the experiments from Boesrma and Pater (2016), as well as Bruce Tesar for making test set available. This research was supported by NSF grants BCS-1650957 and BCS- 2140826 to the University of Massachusetts Amherst.

Proceedings of *AMP 2021*

L indicates that it prefers the Loser.

(1) *Winner-Loser comparisons for the first simple unattested system*

Winner ~ Loser	FtBin	Nonfin	AllFeetR
a. (L L1) ~ (L1)L	W	L	W
b. (H1)L ~ (H L1)		W	L
c. H(H1)L ~ (H1)H L			W

A type of word that gets final stress is shown in (1a.). An unviolated Iambic constraint (not shown here) limits feet to those with stress on the rightmost syllable. The Loser's initial stress thus requires a violation of FtBin, which therefore prefers the Winner with a binary foot. The Winner's satisfaction of FtBin comes at the cost of a violation of Nonfin, which prefers the Loser. When the initial syllable is heavy, as in (1b.), it can be stressed without violating FtBin, and initial stress is preferred over final stress by Nonfin. AllFeetR is violated by the Winner in (1b), but serves to select the Winner in (1c), where Nonfin is not at issue. With the constraints in the order shown (ranked or weighted), the Winners are correctly optimal.

We have been unable to find an example of a language with this pattern of stress in the typological literature that we have surveyed, and no phonologists that we have asked about it could think of one. The only prior discussion of a similar typological gap that we have found is that of Kager (2005, 24), who notes that a pattern in which only disyllables have final stress and all other word types have penultimate is produced by a standard set of Gradient Alignment constraints (McCarthy and Prince 1993), but not by Rhythmic Licensing Theory (Kager 2001), which is a grid-based theory of stress placement. He classifies it as unattested. Kager's typological investigation is limited to quantity insensitive patterns, so the exact pattern here is not considered.

From the perspective of a foot-based theory, it is hard to imagine what the explanation of this gap would be. Each of the constraints producing the pattern has well-attested effects elsewhere, and this pattern of interaction where a lone light syllable is footed with an adjacent one is the iambic version of the analysis that Prince and Smolensky (2004) provide for Latin. It's perhaps worth noting that this suspension of nonfinality in small words is not particular to OT: parametric theories with extrametricality also invoke mechanisms to accomplish it (e.g. Hayes 1995).

This pattern has been previously discussed in the learning literature. Jarosz (2013; fn. 10) notes that none of the OT learners she tested on it using the TS candidate and constraint sets found a correct analysis, and speculates that source of the difficulty is that "learners can be led astray by an overwhelming amount of evidence consistent with a preferred (but incorrect) analysis". A trochaic analysis works for all word types except the light-initial disyllables (these cannot be parsed with final stress without violating the TS version of the Trochee constraint, FootNonfin). The failed learners wound up with a trochaic analysis in Jarosz's study, as well as in our own learning simulations.

Before turning to our own learning results, we'll present another example of a pattern that shares all of these properties: (relatively) simple descriptively, apparently unattested, iambic under a foot-based approach but with most of the data consistent with a trochaic analysis, and as we'll show below, hard to learn. In this one, main stress is always on the penultimate syllable. Secondary stress appears on the initial syllable in words up to four syllables in length, and shifts to the peninitial syllable in longer words. The foot parsings underlying this pattern are shown in (2); a *2* following a syllable indicates secondary stress while a *1* indicates primary. Fixed penultimate primary stress comes from having an iambic foot as close to the right edge as possible, while leaving the final syllable unparsed. Secondary stress appears on the initial syllable when the binary primary stress foot leaves only one syllable for the initial foot (2b.). It shifts to the peninitial syllable when there is more than one syllable to the left of the primary stress foot, as in (2c.).

(2) *Foot parsings for the second unattested language*

a. (L1) L b. (L2) (L L1) L c. (L L2)L (L L1) L

It's again mysterious from the perspective of a foot-based theory of stress why this pattern should be unattested. Each of its parts are well attested: final "invisibility", iambicity, and edge-oriented foot placement.

It can be produced with a standard OT constraint set, including the one TS use, and could also be straightforwardly generated by a parametric theory. Kager (2005) classifies the pattern of penultimate main stress and second syllable secondary as unattested, and points out that it too is generated by Gradient Alignment constraints, but not by Rhythmic Licensing. One can well imagine how it could be hard to learn, given that it has the same property that Jarosz (2013) identified for the previous pattern: much of the data is consistent with a trochaic analysis. Only words six syllables or longer cannot be parsed with trochaic feet, and when our learner fails, it winds up with a trochaic analysis.

3 Hidden structure learning with MaxEnt grammar

In the TS learning problems, prosodic structure is hidden in the sense that the target learning data consist only of strings of heavy and light syllables with stress indicated. A stress pattern for a given type of word will usually be consistent with multiple prosodifications. The learner is supplied with fully prosodified candidates and their constraint violations, and an indication of which candidates are consistent with each learning datum (e.g. that both candidates (L L1)L and L(L1 L) are consistent with the L L1 L learning datum). The goal of learning is to find a ranking of the constraints that will make one of the consistent candidates optimal for every learning datum.

TS propose that the learner parses the learning datum using its current grammar, with the resultant full structure taken as the basis for updating the constraint ranking (Robust Interpretive Parsing; see Jarosz 2013 for alternatives to this proposal). In our MaxEnt approach, first explored in Pater et al. (2012), we use the current probabilities of the consistent candidates to create a violation vector for the update of the constraint weights. The assignment of candidate probabilities by a MaxEnt grammar, and the creation of the violation vector for a learning datum are illustrated in (3). The constraint names are from TS; definitions appear in section 4 below. Constraint violations are shown as negative integers, and the current constraint weights are shown underneath each constraint. The Harmony is the weighted sum of constraint violations, as in Harmonic Grammar (Smolensky and Legendre 2006). In MaxEnt, the probability of a candidate is proportional to the exponential of its harmony; the probability of each of the four candidates is shown in the rightmost column. The probabilities in parentheses for the last two candidates are normalized across just those candidates, rather than the whole candidate set.

The violation vector of the learning datum shown in the bottom row is the probability-weighted sum of violations of the violation vectors of the consistent candidates, that is, of candidates (3c.) and (3d.), using the probabilities normalized over those candidates. The greater penalty assigned by Nonfin and Iambic reflects the greater probability of candidate (3c.), which violates those constraints. Since the violation vector *is* the representation, we can say that the representation assigned to the learning datum is a blend of the representations of the consistent prosodifications, with the strength of each reflecting its probability.

(3) *Violation vector for learning datum L L1 L based on probabilities of consistent prosodifications*

	AllFeetR 3	Nonfin 1	Iambic 1	FootNonfin 1	*Harmony*	*Probability*
a. L(L L1)		–1		–1	–2	0.44
b. (L1 L)L	–1		–1		–4	0.06
c. L(L1 L)		–1	–1		–2	0.44 (0.88)
d. (L L1)L	–1			–1	–4	0.06 (0.12)
L L1 L	–0.12	–0.88	–0.88	–0.12		

The learning objective is to maximize the likelihood of the learning data, or equivalently, to minimize the difference between the probabilities assigned by the grammar and the probabilities in the learning data. To calculate the probability of a learning datum, we sum the probabilities of the consistent candidates. With these constraint weights, the learning datum L L1 L has probability 0.50. Its probability will be maximized as that sum approaches 1, which would happen for example as we raised the weights of AllFeetR and FootNonfin relative to the other constraints, making L(L1 L) more and more probable since it violates neither constraint.

If we assume that the learning data give L L1 L probability 1, the (inferred) observed data is the violation vector in the last row of (3). The learner's expectation is the probability weighted sum over the whole

candidate set, using the probabilities generated by the grammar. To create a vector for the update of the constraint weights we subtract the expected vector from the observed one (in OT terms, the Loser's violations from the Winner's). This calculation is shown in (4).

(4) *Calculation of update vector as Observed – Expected (Winner – Loser)*

	AllFeetR	Nonfin	Iambic	FootNonfin
Observed	–0.12	–0.88	–0.88	–0.12
Expected	–0.12	–0.88	–0.50	–0.50
Obs. – Exp.	0	0	–0.38	+0.38

The update vector is added to the current constraint weights, scaled by a learning rate. This will increase the weight of FootNonfin and decrease the weight of Iambic, thus increasing the probability of the trochaic candidates. Further updates will continue to increase the probability of trochaicity, and will also increase the probability of final feet due to increases of the weight of AllFeetR, and decreases of Nonfin. The learner will therefore assign more and more probability to L(L1 L), getting arbitrarily close to 1 with an increasing number of updates ($p = 1$ is impossible since there will always be some probability assigned to the other candidates). It is important to note that even though it is successful here, this learner, like other hidden structure learners, is not guaranteed to converge on a correct grammar. For a simple example of a local optimum using these constraints that will trap this learner see Boersma and Pater (2016).

We have just described the Gradient Descent algorithm with a form of Expectation Maximization to deal with hidden structure, using a formulation from Staubs and Pater (2013) that emphasizes its relationship to other learning models that have been used in the phonological literature. Run on-line, where each learning update is for a single tableau sampled from the learning data, it would be very similar to using a MaxEnt grammar in Boersma and Pater's (2016) hidden structure learning set-up, which combines TS's Robust Interpretive Parsing with the HG-GLA, whose update rule is equivalent to the one above. The only difference would be that we use probability weighted summation over consistent candidates rather than sampling from the distribution over them. A batch approach also eliminates on-line sampling from the set of tableaux; each learning update is over a probability weighted sum of the whole dataset rather than over a single tableau (see further Staubs and Pater 2013). In fully batch learning that eliminates these two kinds of sampling, which we use here, the outcome for a given set of initial weights is deterministic, which is convenient for learning experiments since it eliminates the need to average over multiple runs.

Given the objective function (maximizing likelihood or minimizing error), there are various possible choices for an algorithm to find weights that optimize it – Gradient Descent is just one. We follow Pater et al. (2012) in choosing L-BFGS-B (Byrd et al. 1995), which has the advantage over other optimization algorithms of allowing a minimum weight of zero to be imposed (negative weights make constraints have the opposite of their intended effect). Because it is significantly more complex than Gradient Descent, we refer the interested reader to Byrd et al. (1995) for a description, and note just that the approach to hidden structure and to calculating the observed and expected probabilities are the ones explained here.

4 Learning experiments

The TS test set consists of 124 languages that can be generated by a ranking of the 12 constraints in (5).

(5) *Constraints from the TS Test Set* (constraint definitions from Jarosz, 2013)

1. *FtBin*: Each foot must be either bimoraic or disyllabic.
2. *Parse*: Each syllable must be footed.
3. *Iambic*: The final syllable of a foot must be the head.
4. *FootNonfin*: A head syllable must not be final in its foot.
5. *Nonfin*: The final syllable of a word must not be footed.
6. *WSP*: Each heavy syllable must be stressed.
7. *WordFootR*: Align the right edge of the word with a foot.
8. *WordFootL*: Align the left edge of the word with a foot.
9. *MainR*: Align the head foot with the right edge of the word.

10. *MainL*: Align the head foot with the left edge of the word.
11. *AllFeetR*: Align each foot with the right edge of the word.
12. *AllFeetL*: Align each foot with left edge of the word.

There are 62 tableaux, one for every combination of light and heavy syllables for words two syllables through five syllables in length, and one each for six and seven syllable strings of light syllables. Candidates are all the possible footings of those strings into maximally binary feet, with one of the feet being the primary stress. The learning datum for each tableau is a stress pattern over the string of light and heavy syllables, that is, an indication of the correct placement for the primary stress and any secondary stresses.

For each of the 124 languages we ran our learner[1] with all weights starting at 1, and also 10 times with each weight randomly sampled from a uniform distribution from 0 to 10. We applied two success criteria, which turned out to be equivalent in that they agreed on whether each run was successful (see Pater 2014 on success criteria for this type of learning problem). In one, a learning run was successful if in every tableau the candidate with the highest probability was consistent with the learning datum. In the other, a learning run was successful if the summed probability of the consistent candidates was > 0.90. When the learner is initialized with constraint weights set to 1, 115/124 of the languages are learned successfully (93%). When the initial weights were randomly sampled, the learner was successful 91% of the time across 10 runs of each of the languages. The randomly initialized learner was successful in at least one of the ten runs for 121/124 (98%) of the languages. Two of the languages that were never successfully learned with random initialization were learned successfully with initialization at 1, leaving just one language that our learner never successfully learned (labeled language 79 in the test set).

We also ran learning experiments using the same setup as in Boersma and Pater (2016), using a Praat script and input files supplied by Paul Boersma. The learners in this case are on-line, and always start with random rankings or weight initializations; see Boersma and Pater (2016) for further details. We replicated the Stochastic OT and Noisy HG experiments from that paper, and also ran new experiments with a MaxEnt grammar model, and Noisy MaxEnt, which adds to MaxEnt the evaluation noise of Stochastic OT and Noisy HG. We ran each learner 10 times with random initializations. The success rates calculated as an overall average over runs, and in terms of whether each language was learned successfully at least once, are shown in (6), alongside the success rates for our batch MaxEnt learner. Our results are better than any of these on-line learners, and just slightly worse than Jarosz's (2013) best results of a 95% overall average with an on-line OT and HG learners using an alternative parsing strategy to that of TS and Boersma and Pater (2016), and Jarosz's (2015) 96% overall average using her Expectation Driven Learner with a pairwise ranking grammar. Language 79 was also never successfully learned by any of the on-line learners we ran.

(6) *Overall success rates for batch MaxEnt and on-line learners*

	Batch MaxEnt (init 1)	Batch MaxEnt (random init)	On-line MaxEnt	Noisy MaxEnt	Stochastic OT	Noisy HG
Average	93%	91%	83%	90%	60%	89%
At least one	93%	98%	92%	94%	70%	90%

Using the results of our batch learner, we identified 11 languages as "Hard": the nine languages that initialization at 1 failed on, and two more that random initialization never succeeded on. Of those 11 languages, all of which seem unattested, 8 are cases in which a correct analysis must use iambic feet with nonfinality to place primary stress on the penultimate syllable. The other 3 languages are ones that have complex patterns of quantity sensitive stress placement in which a correct analysis uses a mix of trochaic and iambic feet. Language 79, the hardest of the languages, is in this second group. We were able, with some effort, to construct an analysis by hand for language 79, but it remains unclear what distinguishes it and the other two languages in this group from some other languages in the TS test set that our learners find relatively easy. We therefore put these aside and focus on the iambic nonfinality languages.

[1] A Python implementation of our learner, with input files for the TS languages, and output files for the runs reported here, is available at https://github.com/blprickett/Hidden-Structure-MaxEnt.

6 Successful and failed analyses of iambic nonfinality

The first language that we discussed in section 2 is language 52 of the TS test set. The version of the test set we used (from Bruce Tesar via Paul Boersma) does not come with correct rankings or descriptions of the languages, so the descriptions and analyses presented here are our own. We start with a set of weights that we found by hand that yields a correct analysis, that is, one in which a candidate that is consistent with the target learning datum has the highest probability (=highest Harmony) for all of the tableaux. The constraints with non-zero weights are shown in (7), alongside the parse of the bisyllabic word types and one trisyllable. This analysis adds Parse and Iambic to the constraint set and ordering discussed in section 2. Iambic is unviolated in this analysis and could also have an arbitrarily higher weight; Parse is violated in order to satisfy Nonfin and must therefore have a lower weight than that constraint. The analysis will fail if either of these constraints has zero weight (e.g. by failing to choose iambs over trochees).

(7) *Weights and parses for correct analysis of language 52 found by hand*

FtBin	15	(L L1)
Nonfin	10	(L H1)
AllFeetR	5	(H1) L
Parse	1	(H1) H
Iambic	1	(L L1) L

Our learner, with a random initialization, found the weights in (8), rounded to one decimal point, which yield the same parses as shown in (7). In this analysis, Nonfin has the highest weight, and the fact that (L L1) and (L H1) have higher probability than (L1)L and (L1)H respectively comes from a gang effect. FtBin, AllFeetR, MainR, WordFootR and Parse all prefer final stress for both L L and L H, as does WSP for L H, and their summed weight is greater than that of Nonfin.

(8) *Correct weights for language 52 found with a batch MaxEnt learner (random init)*

Nonfin	29.3
FtBin	17.3
Iambic	16.1
AllFeetR	11.5
MainR	5.1
AllFeetL	3.4
WordFootR	2.6
FootNonfin	2.2
Parse	1.3
MainL	0.9
WSP	0.8

Random initialization led to successful learning only one out of ten tries for this language, and initialization at 1 also failed, as did every run of the Praat on-line learners. The final weights from initialization at 1 are shown in (9), along with the parses to which they assign highest probability for the five word types also shown in (7). The asterisked word types have stress in the wrong place relative to the learning data. This is a trochaic analysis, with FootNonfin, the TS trochee constraint, having higher weight than Iambic, and constraints wanting a foot at the right edge having higher weight than Nonfin, which had zero weight.

(9) *Incorrect weights for language 52 found with a batch MaxEnt learner (init 1)*

AllFeetR	8.5	*(L1 L)
WordFootR	6.3	*(L1 H)
FootNonfin	5.6	(H1 L)
MainR	4.4	(H1 H)
AllFeetL	2.2	L (L 1L)
Iambic	1.9	

MainL	1.8
WSP	1.2
Parse	1.1

As Jarosz (2013) points out, the trochaic analysis fails on only 2 of the 62 tableaux. For the learners to be winding up so often in this local optimum, there must be something that prefers the trochaic analysis of penultimate stress over the iambic one. One likely factor is the greater depth of constraint ordering in the iambic analysis, which requires an active AllFeetR constraint that is itself overridden by Nonfin. The placement of the foot in the trochaic analysis is supported by several constraints: AllFeetR, WordFootR, and MainR, all of which have relatively high weight in (9). In MaxEnt, the iambic analysis requires AllFeetR to have sufficient weight on its own to rule out feet further to the left, and Nonfin must have sufficiently higher weight than AllFeetR to keep the foot off the final syllable. The trochaic analysis allows a group of lower weighted constraints to jointly place the foot in final position. This seems likely to structure the learning space in such a way that it is easier to wind up with the trochaic analysis, and harder to abandon it, though this remains somewhat speculative.

The second language in section 2 is language 121 in the TS test set. One of the randomly initialized batch MaxEnt learners found the correct iambic analysis in (10). Nonfin has the highest weight and is unviolated in the highest probability candidates, and Iambic has higher weight than FootNonfin (not shown due to its zero weight). This language differs from language 52 in having an initial secondary stress foot, which emerges because the constraints that prefer it, Parse and WordFootL, have higher weight than AllFeetR, which assigns violations to feet not aligned with the right edge.

(10) *Correct weights for language 121 found with a batch MaxEnt learner (random init)*

Nonfinal	42.6	(L1) L
MainR	19.9	(L2) (L L1) L
Iambic	13.5	(L L2) (L L1) L
AllFeetL	12.8	(L L2)L (L L1) L
Parse	10.4	
WordFootL	8.0	
AllFeetR	2.0	
MainL	0.7	
FtBin	0.5	

Intialization at 1 failed to yield a correct analysis, as did 8/10 of the random initializations with the batch MaxEnt learner. The on-line Praat learners were relatively successful on this language, with the exception of the Stochastic OT learner, which always failed. The failed trochaic analysis produced by the final weights of the batch MaxEnt learner with initialization at 1 is shown in (11). FootNonfin chooses trochaic bisyllabic feet, and the right oriented foot placement constraints places the main stressed foot at the right edge. The secondary stress foot is also aligned to the right, which works for all word types except the 6 and 7 syllable words, of which there are only two in the test set. The lack of an initial foot in the six syllable word is due to the activity of AllFeetR, which would assign a violation score of –4 to the initial foot, resulting in a penalty of 18 that exceeds the gain on Parse of 17 (2*8.5) that the foot would provide (see Potts et al. 2010 on OT/HG differences with Alignment constraints). If a trochaic foot were there, the result would still be incorrect, since the first and third syllables would be stressed, rather than the second.

(11) *Incorrect weights for language 121 found with a batch MaxEnt learner (init 1)*

FootNonfin	26.8	(L1 L)
MainR	10.6	(L2 L) (L1 L)
Parse	8.5	L (L2 L) (L1 L)
AllFeetR	4.5	*L L (L2 L) (L1 L)
WordFootR	2.9	
FtBin	2.2	
AllFeetL	2.0	

The other languages with iambic nonfinality amongst our "hard" languages are languages 41, 44, 45, 56 and 119 in the test set. These have more data that are inconsistent with a trochaic analysis, and were generally somewhat more frequently successfully learned in our experiments. The proportions of successful runs over all 11 batch MaxEnt runs per language (initialization at 1, and 10 random), and over all 40 runs of on-line learners (10 each for the four grammar models) are provided in (12). The last two rows provide the average for these 7 languages, and for the 124 in the test set. Recall that a failure on initialization at 1 was sufficient for our "hard" classification, so it would be possible for a hard language to have a success rate as high as the average over all 124 (i.e. 10/11 = 0.91).

(12)

Language	*Batch MaxEnt*	*On-line*
41	0.73	0.80
44	0.36	0.30
45	0.45	0.68
52	0.09	0.00
56	0.45	0.35
119	0.18	0.00
121	0.18	0.70
All 7	0.35	0.40
All 124	0.90	0.81

We note that language 119 was no easier than 121 for the batch MaxEnt learner, and was in fact harder for the on-line learners we explored, even though less of the data for 119 is consistent with a trochaic analysis. Language 119 differs from 121 only in the four syllable words, which have stress on the second and third syllables, rather than the first and third. As shown in (11), the four syllable pattern in language 119 can be parsed trochaically; the second + third pattern in language 121 cannot. We do not know why language 119 was not easier than 121. All the other languages in which more data was inconsistent with trochaic analysis (41, 44, 45, 56) were more successfully learned than the two languages that we have discussed in detail in which only two of the 62 learning data were inconsistent with trochaicity (51 and 121).

7 Conclusions and future directions

Boersma (2003) first brought up the possibility that the failure of learners on parts of the TS test set might help to explain typological gaps, if those languages were unattested. This possibility has remained unexplored until now, but the initial results presented in this paper seem promising. We implemented a batch MaxEnt learner that uses a form of Expectation Maximization to deal with hidden structure, and examined its performance on the TS test set. We found that the 11 languages that we classified as "hard" for that learner do seem to be unattested, and that 8 of them share a common property: penultimate stress assigned by an iambic foot that is kept off the final syllable by Nonfinality. Learners of these 8 languages often fall into local optima in which they have a trochaic analysis of penultimate stress, and when this analysis fails only for a small portion of the dataset, the probability of finding a correct analysis becomes quite low.

We provided reasons that the trochaic analysis might be preferred for penultimate stress, but did not provide any detailed analysis of the structure of the learning space and of the local optima within it. We have done some initial exploration of smaller learning problems in which we have been able to study the interacting effects of the weight initialization, the distribution over the learning data, and the structure of the constraint set. Our hope is that this work will be able to scale up to larger systems like the ones we studied here.

This research extends earlier demonstrations that learning might shape stress typology (Bane and Riggle 2008; Heinz 2009; Staubs 2014; Stanton 2016) to cases of ambiguity caused by hidden structure. Further work is needed to show that learning will in fact have the desired effect on typology (e.g. through agent-based simulations – see Staubs 2014), and it also remains possible that the true source of the gaps is in the grammatical theory: the systems are not produced by metrical theories that place stress without reference to feet (Prince 1983; Bailey 1995; Gordon 2002; Kager 2005).

As well as further exploring the tightening of metrical typology through the use of learning, there is a great deal of work to be done on broadening the empirical coverage of combined learning-grammar systems. As a first step in studying a broader typology, we created a test set using the TS constraints and candidate

sets for the patterns in the StressTyp2 database (Goedemans, Heinz, and Hulst 2015) for which Finite State Automata are supplied, which we used to generate the target stress patterns for each language. We eliminated languages that used a three-way weight distinction, which cannot be captured by the TS constraint set. Our learner, with an initialization of 1, succeeded on just 45% of them. Inspection of the languages on which the learner fails suggests that many – if not all – are in fact patterns that cannot be expressed with the TS constraints. One example is the well-known Latin main stress pattern: stress the penult if heavy, else the antepenult. Because TS use a version of the Trochee constraint that is violated by a monosyllabic foot (FootNonfin), Prince and Smolensky's (2004) analysis of this pattern is unavailable. When we add a more standard Trochee constraint that penalizes only finally stressed disyllables, our learner does succeed on the pattern. The questions of what is needed for a constraint set to generate the full typology represented in StressTyp2 and in other sources, and whether a learner equipped with such a constraint set would succeed on all the patterns, are topics that we hope to address in further work.

References

Bailey, Todd M. 1995. "Nonmetrical Constraints on Stress." PhD Thesis, 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*, 10:29–38. Association for Computational Linguistics.

Boersma, Paul. 2003. "Review of Tesar & Smolensky (2000): Learnability in Optimality Theory." *Phonology* 20: 436–46.

Boersma, Paul, and Joe Pater. 2016. "Convergence Properties of a Gradual Learner in Harmonic Grammar." In *Harmonic Grammar and Harmonic Serialism*, edited by John J. McCarthy and Joe Pater, 389–434. Bristol, Connecticut: Equinox Publishing.

Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. "A Limited Memory Algorithm for Bound Constrained Optimization." *SIAM Journal on Scientific Computing* 16 (5): 1190–1208. https://doi.org/10.1137/0916069.

Goedemans, Rob, Jeffrey Heinz, and Harry van der Hulst. 2015. *StressTyp2, Version 1*. http://st2.ullet.net.

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*, edited by Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–20.

Gordon, Matthew. 2002. "A Factorial Typology of Quantity-Insensitive Stress." *Natural Language & Linguistic Theory* 20 (3): 491–552.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press.

Heinz, Jeffrey. 2009. "On the Role of Locality in Learning Stress Patterns." *Phonology* 26 (2): 303–51.

Jarosz, Gaja. 2006. "Rich Lexicons and Restrictive Grammars: Maximum Likelihood Learning in Optimality Theory." PhD Thesis, Johns Hopkins University. https://rucore.libraries.rutgers.edu/rutgers-lib/37916/.

———. 2013. "Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing." *Phonology* 30 (01): 27–71.

———. 2015. "Expectation Driven Learning of Phonology." University of Massachusetts Amherst. http://blogs.umass.edu/jarosz/2015/08/24/expectation-driven-learning-of-phonology/.

Kager, René. 2001. "Rhythmic Directionality by Positional Licensing." Handout from Fifth HIL Phonology Conference (HILP 5), University of Potsdam, 11 January 2001. http://roa.rutgers.edu/article/view/524.

———. 2005. "Rhythmic Licensing Theory: An Extended Typology." In *Proceedings of the Third International Conference on Phonology*, 5–31. Seoul National University.

McCarthy, John J., and Alan M. Prince. 1993. "Generalized Alignment." In *Yearbook of Morphology 1993*, edited by Geert Booij and Jaap van Marle, 79–153. Kluwer.

Pater, Joe. 2014. "Categorical Correctness in MaxEnt Hidden Structure Learning." *UMass Computational Phonology* (blog). September 24, 2014. https://blogs.umass.edu/comphon/2014/09/24/success-maxent/.

Pater, Joe, Robert Staubs, Karen Jesney, and Brian Smith. 2012. "Learning Probabilities over Underlying Representations." In *Proceedings of the Twelfth Meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology*, 62–71.

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 (1): 77–117.

Prince, Alan. 1983. "Relating to the Grid." *Linguistic Inquiry* 14: 19–100.

———. 2002. "Arguing Optimality." In *Papers in Optimality Theory II*, edited by Andries Coetzee, Angela C. Carpenter, and Paul de Lacy. Amherst, MA: Graduate Linguistics Society Association, UMass Amherst. http://ruccs.rutgers.edu/images/personal-alan-prince/gamma/argopt.pdf.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Smolensky, Paul, and Geraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, Massachusetts: MIT Press.

Stanton, Juliet. 2016. "Learnability Shapes Typology: The Case of the Midpoint Pathology." *Language* 92 (4): 753–91. https://doi.org/10.1353/lan.2016.0071.

Staubs, Robert. 2014. "Computational Modeling of Learning Biases in Stress Typology." Doctoral dissertation, University of Massachusetts Amherst. http://scholarworks.umass.edu/dissertations_2/230/.

Staubs, Robert, and Joe Pater. 2013. "Modeling Learning Trajectories with Batch Gradient Descent." MIT, Cambridge MA. http://people.umass.edu/pater/pater-staubs-gradient-descent-2013.pdf.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. The MIT Press.