Learning exceptionality and variation with lexically scaled MaxEnt[*]

Coral Hughto, Andrew Lamont, Brandon Prickett, and **Gaja Jarosz**
University of Massachusetts Amherst
{coralwilliam, alamont, bprickett, jarosz}@linguist.umass.edu

Abstract

A growing body of research in phonology addresses the representation and learning of variable processes and exceptional, lexically conditioned processes. Linzen et al. (2013) present a MaxEnt model with additive lexical scales to account for data exhibiting both variation and exceptionality. In this paper, we implement a learning model for lexically scaled MaxEnt grammars which we show to be successful across a range of data containing patterns of variation and exceptionality. We also explore how the model's parameters and the rate of exceptionality in the data influence its performance and predictions for novel forms.

1 Introduction

While phonological research often focuses on categorical generalizations, a growing body of research addresses the representation and learning of variable processes and exceptional processes, where application is lexically conditioned (see Coetzee and Pater (2011) and Pater (2010) for overviews). A few recent studies have modeled processes that exhibit both variation *and* exceptionality (Hayes and Londe, 2006; Pater et al., 2012; Linzen et al., 2013; Nazarov, 2018; Shih, 2018; Zymet, 2018).

Linzen et al. (2013) model co-existing exceptionality and variation in Russian using a Maximum Entropy (MaxEnt) grammar (Goldwater and Johnson, 2003) with additive, lexically specified scales. Russian contains a vowel alternation process that exhibits both variation and idiosyncratic lexical conditioning (exceptionality). Linzen et al. show that speakers apply this process variably and that its variation differs across lexical items. In their lexical scaling framework, each lexical item is associated with a vector of scales that are added to the general weights of the grammar's constraints. These summed weights are used to calculate the probability of the input's surface realization. This allows the likelihood of a phonological process to differ across morphemes, since the scales can modulate how constraints are weighted for different lexemes. While Linzen et al. (2013) show that a lexically scaled MaxEnt grammar can successfully represent Russian speakers' knowledge of a pattern that is both variable and exceptional, they do not show how such a grammar would be learned.

In this paper, we introduce a model for learning lexically scaled MaxEnt grammars from data exhibiting both variation and exceptionality.[1] The primary challenge for formalizing learning in this framework is generalizing appropriately beyond the learning data and limiting the learner's reliance on lexical scales. Since every morpheme can potentially scale the weight of every constraint, there is potential for massively over-fitting the learning data

[*] We wish to thank the SCiL 2019 anonymous reviewers, the UMass Linguistics Sound Workshop, Brendan O'Connor, and Joe Pater for valuable comments.

[1] Our code is publicly available at https://github.com/chughto/Lexically-Scaled-MaxEnt

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 91-101.
New York City, New York, January 3-6, 2019

and failing to generalize. We approach this challenge as a problem of feature selection and seek a learner that utilizes scales (i.e., assigning them nonzero weights) only when needed to account for lexical conditioning. We propose an objective function relying on an L1 (linear) prior (§2), rather than the more commonly used L2 (quadratic) prior (§4.1), to formalize these criteria.

Our approach differs in a number of ways from previous models for learning exceptionality and variation. We assume the learner must induce a weighting for general phonological constraints and make lexical conditioning choices without prior knowledge of which lexical items behave exceptionally (Allen and Becker, 2015; Becker and Gouskova, 2016). Rather than splitting the learning of general phonological patterns and the learning of exceptions/classes into distinct learning phases (Nazarov, 2018; Shih, 2018), or treating the learning of lexical conditioning as emergent from repeated exposure to the lexicon (Zuraw, 2000; Zuraw, 2010), we seek to formally characterize the criteria that favor the desired balance of lexical sensitivity and generalization in a model that optimizes general weights and lexical conditioning in parallel. Our approach is most similar to Moore-Cantwell and Pater (2016); however, we argue for an L1 prior rather than an L2 prior (§4.1). We demonstrate the capacity of our model to learn variation and exceptionality using a variety of toy languages based on the Russian process mentioned above (§3).

We also explore the model's predictions for novel data, examining how the learner decides which patterns to treat as exceptional and which to generalize (§4). Previous behavioral investigations of speakers' productive knowledge of lexically conditioned (morpho-)phonological alternations have found that speakers extend statistical tendencies in the lexicon to novel forms (Zuraw, 2000; Ernestus and Baayen, 2003; Hayes and Londe, 2006; Hayes et al., 2009; Linzen et al., 2013; Becker and Gouskova, 2016). In some cases, the absolute rates of application of a process in nonce forms closely follow rates observed in the lexicon, yielding so-called "frequency-matching" behavior (Hayes and Londe, 2006; Hayes et al., 2009; Zymet, 2018). In other cases, however, rates of application of exceptional processes are systematically skewed lower as compared to the lexical rates (Zuraw, 2000; Albright and Hayes, 2003; Ernestus and Baayen, 2003). Under a variety of learning assumptions, frequency-matching behavior is not automatic. To better understand some of the factors that may play a role in these divergent findings, we examine the properties of the data distribution and parameters of the model that affect frequency-matching behavior on nonce forms.

2 Lexically scaled MaxEnt

Linzen et al.'s (2013) scaled weights framework uses weighted constraints to represent probabilistic phonological patterns, and adds scales on those weights to represent lexicalized behavior for individual morphemes. In MaxEnt (Goldwater and Johnson, 2003), the probability of some surface representation (SR), given a grammar and an underlying representation (UR), is calculated as in (1):

$$p(i) = \frac{e^{\mathscr{H}_i}}{\sum_{k \in K_i} e^{\mathscr{H}_k}} \tag{1}$$

Here $\mathscr{H}_i$ is the harmony of a given (UR, SR) pair i, and K_i is the set of candidates that share the same UR as i (including i itself). Typically, harmony is the weighted sum of a candidate's constraint violations (Goldwater and Johnson, 2003), however in Linzen et al.'s (2013) framework, harmony is a function of a candidate's violations, the general weights, and the relevant scales, as shown formally in (2):

$$\mathscr{H}_i = \sum_{\gamma \in \Gamma} (w_\gamma + \sum_{m \in \mu_i} s_{\gamma m})(v_{\gamma i}) \tag{2}$$

Here Γ is the set of constraints, w_γ is the general weight of constraint γ, μ_i is the set of morphemes in the (UR, SR) pair i, $s_{\gamma m}$ is the scale that morpheme m has for constraint γ, and $v_{\gamma i}$ is the number of violations assigned to candidate i by constraint γ. The parameters of the model are the general constraint weights and the additive lexical scales. For succinctness, we refer to these simply as "weights" and "scales", respectively. Every morpheme is associated with a scale for every constraint that is added to that constraint's weight. This model is closely related to other approaches relying on additive scales (Boersma and Hayes, 2001; Coetzee and Kawahara, 2013; Hsu and Jesney, 2016) and multiplicative scales (Kimper, 2011). However, in Linzen et al.'s framework, scaling is not restricted to faithfulness constraints or systematic factors (e.g., register, frequency): all constraints are available for lexical scaling by all morphemes. Using both weights and scales enables the model to represent lexicalized exceptions by employing the scales to modulate the effect of some constraints. We leave exploring the relationship between this approach and indexed constraints (Kraska-Szlenk, 1995; Pater, 1996), a closely related framework, to future work.

We formalized learning as minimizing the objective function in (3), the sum of the negative log likelihood and an L1 prior on weights and scales.

$$-\sum_{i} \log p(i) + C \sum_{\gamma \in \Gamma} |w_\gamma| + C \sum_{\gamma \in \Gamma} \sum_{m \in M} |s_{\gamma m}| \quad (3)$$

Here, M is the set of morphemes in the language, and C is a parameter that controls the overall strength of the prior. Our goal was to determine whether learning of phonological generalizations could occur without formally distinguishing between weights and scales. Accordingly, this prior penalizes both weights and scales with a single strength parameter C. In simulations reported here, both weights and scales are restricted to nonnegative values, but this is not a inherent restriction of the model. For optimization, we used a form of gradient descent adapted for L1 priors – the "L1 (Clipping)" method described by Tsuruoka et al. (2009).

3 Learning variation and exceptionality

To explore the capacity of this model to learn a range of variable and exceptional patterns, it was trained on four toy languages based on the Russian vowel alternation described by Linzen et al. (2013).

In Russian, underlyingly CV prepositions surface as C before words beginning with vowels or single consonants; we follow Linzen et al. (2013, §5.1) in treating this alternation as deletion. Before words beginning with consonant clusters, vowel deletion is variable and lexically conditioned. For example, the vowel in /sa/ "from, with" variably surfaces with certain cluster-initial words (4a), categorically deletes with certain others (4b), and categorically surfaces with others (4c) (Linzen et al., 2013, 455).

(4) a. [s ∼ sa] mnóʐəstvəm "with a large amount, (mathematical) set"
b. [s ∼ *sə] prikázəm "with the order"
c. [*s ∼ sə] stərikóm "with the old man"

The factors influencing vowel deletion in Russian span multiple phonological dimensions such as stress and sonority profile. For the purposes of testing our learning model, we focused on whether words began with one or two consonants. The four toy languages consisted of 3 prefixes /ape-/, /ate-/, and /ake-/ concatenated with 420 stems, giving 1260 forms in total. Stems were all consonant-initial, beginning either with a single consonant ("C-stems"), or a biconsonantal cluster ("CC-stems"). Six consonants were used {v, r, l, n, s, t}, giving 6 unique C-stem types, and 36 CC-stem types. Each stem type was replicated 10 times, yielding 420 stems in total.

In all four languages, prefix vowels categorically deleted with C-stems, e.g., /ape-naba/ → [apnaba]. Vowel deletion was conditioned with CC-stems, either categorically failing to apply (§3.1) or with its application subject to free variation (§3.2), lexical specification (§3.3), or both (§3.4).

We used three categorically evaluated constraints (see Linzen et al. (2013, 489-490)): ALIGN, MAX, and *CCC. ALIGN prefers vowel deletion, and is violated by candidates containing the final prefix vowel. MAX disprefers vowel deletion, and is violated by candidates lacking the final prefix vowel. *CCC is violated by candidates with triconsonantal clusters, and so disprefers deletion with CC-stems.

After training, the model was tested by evaluating its performance on the learning data and its predictions on a set of nonce forms comprising 3 novel prefixes concatenated with 42 novel stems. Following previous work, we assume that predictions for novel forms are generated using only the general weights.

Learning was evaluated according to quantitative and qualitative criteria. Quantitatively, learning was considered successful if the KL-Divergence (Kullback and Leibler, 1951) between the likelihood assigned by the model and the observed probability in the training data was close to zero, indicating that the model succeeded in accounting for the learning data.[2] Qualitatively, learning was considered successful only if the model appropriately divided weight between the general constraints and

[2] MaxEnt grammars cannot exactly represent categorical behavior, but probabilities can get arbitrarily close to 0 or 1.

	*CCC	Max	Align		
/ape-taba/	11.5	0.0	4.5	*O*	*E*
a. apetaba	0	0	-1	0.00	0.00
b. aptaba	0	-1	0	1.00	1.00
/ape-tnaba/	11.5	0.0	4.5	*O*	*E*
a. apetnaba	0	0	-1	1.00	1.00
b. aptnaba	-1	-1	0	0.00	0.00

Tableau 1: Categorical language

the scales, only using scales when presented with lexically conditioned data. Finally, we required that the model generalizes the observed pattern to nonce forms, deleting (nearly) categorically for novel C-stems while predicting variation for CC-stems in languages with variation and/or lexical conditioning.

For all experiments in this section, the model was run with weights and scales initialized at 0.0, for 20,000 epochs, with a learning rate of .001, and prior term C set to 1.0, unless otherwise noted.

Overall, the model performed well, successfully learning the four toy languages and using the scales appropriately. In all runs, Align received non-zero weight, reflecting (near) categorical vowel deletion with C-stems. In languages with variable or exceptional deletion, *CCC was weighted closer to Align, predicting variation in nonce forms.

3.1 Categorical language

In the Categorical language, prefix vowels always delete with C-stems and never with CC-stems. The solution learned by the model captures this pattern using the general weights only, putting no weight on the scales, as summarized in Table (1). The weight on *CCC is much higher than the weight on Align so that tri-consonantal clusters block vowel deletion, and the weight of Align is above that of Max, so that prefix vowels always delete with C-stems.

	*CCC	Max	Align
General Weights	11.5	0.0	4.5
Morpheme Scales	0.0	0.0	0.0

Table 1: Categorical weights and mean scales

The model's performance on forms in the training data is illustrated in Tableau (1) with a C-stem, /taba/, and a CC-stem, /tnaba/. Candidate probabilities observed in the training data are given in column O. Column E gives the expected candidate probabilities generated by the model, rounded to two decimal places. The model fits the training data extremely well (KL divergence ≈ 0.002) and, because only the general weights are used, the model predicts that the trained pattern should generalize, yielding the same predicted probabilities for nonce forms.

3.2 Variable language

In the Variable language, prefix vowels always delete with C-stems and variably delete 33% of the time with CC-stems. As desired, the model captures this pattern using the general weights only, putting no weight on the scales, as shown in Table (2).

	*CCC	Max	Align
General Weights	5.2	0.0	4.5
Morpheme Scales	0.0	0.0	0.0

Table 2: Variable weights and mean scales

The model's performance on trained forms is illustrated in Tableau (2) below with a C-stem and a CC-stem. The weight of *CCC is above that of Align, but by a smaller margin than in the Categorical language, yielding variable rather than categorical deletion with CC-stems. Max is weighted below Align, so that deletion occurs (nearly) categorically for C-stems. The probabilities generated by the model (E) fit the training data (O) extremely well (KL divergence ≈ 0.002) and, because only the general constraints are used, the model predicts that the trained pattern should generalize, yielding the same predicted probabilities for nonce forms.

	*CCC	Max	Align		
/ape-taba/	5.2	0.0	4.5	*O*	*E*
a. apetaba	0	0	-1	0.00	0.01
b. aptaba	0	-1	0	1.00	0.99
/ape-tnaba/	5.2	0.0	4.5	*O*	*E*
a. apetnaba	0	0	-1	0.67	0.67
b. aptnaba	-1	-1	0	0.33	0.33

Tableau 2: Variable language

3.3 Lexical language

The Lexical language is identical to the Categorical language, except that one prefix, /ape-/, is exceptional: its vowel always deletes with CC-stems. Averaged across the lexicon, the rate of deletion with CC-stems is therefore 33%, but this pattern cannot be captured using the general weights alone. The scales must be used to distinguish the behavior of the exceptionally deleting prefix from the other two prefixes. As Table (3) shows, the model's solution weights the general constraints in the same order as in the Categorical language: *CCC $>$ ALIGN $>$ MAX. The model additionally scales up the weight of ALIGN for the deleting prefix /ape-/, yielding (near) categorical deletion for it, and scales up the weight of *CCC for each of the non-deleting prefixes, preventing deletion for those prefixes.

	*CCC	MAX	ALIGN
General Weights	4.6	0.0	4.1
Deleting Prefix	0.0	0.0	6.4
Non-Deleting Prefixes	5.3	0.0	0.0
Stems	0.0	0.0	0.0

Table 3: Lexical weights and mean scales

The model's performance on forms in the training data is illustrated in Tableau (3) with a CC-stem /tnaba/ paired with a non-deleting prefix /ake-/, and the deleting prefix /ape-/. The weights shown for each input are the sums of the general weights and scales associated with the input morphemes for each constraint. The model's solution fits the training data well (KL divergence ≈ 0.004), predicting deletion for the deleting prefix and no deletion for each non-deleting prefix. These weights additionally yield deletion of all prefix vowels before C-stems (not shown) in the learning data, as expected. Since this is captured by general constraint weights, the same prediction is made for novel C-stems.

Because the general weight of ALIGN is somewhat lower but still close to the general weight of *CCC, variable deletion (37%) is predicted for novel prefixes attached to novel CC-stems. Tableau (4) illustrates with the nonce prefix /aʔe-/ and the nonce stem /pmaba/. Because this form was not present in the training data, only the expected probabilities are reported. As discussed above, predicting variable deletion is desirable given experimental findings that speakers extend lexical trends to nonce forms (Hayes et al., 2009). Deletion is the dispreferred outcome in both the training data and the predictions for novel forms, but the predicted rate of deletion for novel forms (37%) is a little higher than that observed in the training data (33%).

	*CCC	MAX	ALIGN		
/ake-tnaba/	10.0	0.0	4.1	*O*	*E*
a. aketnaba	0	0	-1	1.00	1.00
b. aktnaba	-1	-1	0	0.00	0.00
/ape-tnaba/	4.6	0.0	10.5	*O*	*E*
a. apetnaba	0	0	-1	0.00	0.00
b. aptnaba	-1	-1	0	1.00	1.00

Tableau 3: Lexical language – known prefixes and stems; /ape-/ exceptionally undergoes vowel deletion with CC-stems

Interestingly, by using scales for each prefix, the model did not single out any prefix as qualitatively exceptional, despite the fact that such a solution is available. Removing the weight from the scales of the non-deleting prefixes and dividing it between the weight of general *CCC and the deleting prefix's scale of ALIGN produces a solution that is identical in terms of fit to the training data and the total sum of weights across all constraints and scales. That solution identifies only the deleting prefix as exceptional, and produces different predictions for nonce forms. The proposed objective function does not always differentiate among distinct ways of encoding exceptionality. The solution selected by the model in this experiment is arbitrarily influenced by starting the weights at zero. In experiments with weights initialized to random values between 0 and 10, the solutions selected by the model all have equivalent fit to the training data and total weight but vary somewhat in terms of how exceptionality is encoded and the deletion rate predicted for nonce forms. The availability of such varied solutions depends on the rate

	*CCC	MAX	ALIGN	
/aʔe-pmaba/	4.6	0.0	4.1	*E*
a. aʔepmaba	0	0	-1	0.63
b. aʔpmaba	-1	-1	0	0.37

Tableau 4: Lexical language – nonce prefix and stem

of exceptionality in the training data and the prior. These factors are further explored in §4.

3.4 Variable-Lexical language

The Variable-Lexical language is largely identical to the Variable language, except that 20% of CC-stems are exceptional triggers of categorical vowel deletion. The model's solution is summarized in Table (4). Tableau (5) illustrates the learned weights with a triggering stem /vraba/ and a non-triggering stem /tnaba/. The model learned a set of general weights which closely, but not exactly, reproduces the trained general pattern of variable deletion, and weights the scale of ALIGN higher for triggering CC-stems, though not enough to yield (near) categorical deletion. Again, ALIGN is weighted sufficiently above MAX to motivate (near) categorical deletion with C-stems. The model's fit to the training data for the Variable-Lexical language, while worse than the other languages, is still good (KL divergence ≈ 0.08). Examining the general weights, the model generalizes appropriately, predicting (near) categorical deletion for novel C-stems, and variable deletion for novel CC-stems.

	*CCC	MAX	ALIGN
General Weights	4.8	0.0	4.5
Prefixes	0.0	0.0	0.0
Exceptional Stems	0.0	0.0	1.0
Regular Stems	0.0	0.0	0.0

Table 4: Variable-Lexical weights and mean scales

/ape-vraba/	*CCC 4.8	MAX 0.0	ALIGN 5.5	O	E
a. apevraba	0	0	-1	0.00	0.33
b. apvraba	-1	-1	0	1.00	0.67
/ape-tnaba/	4.8	0.0	4.5	O	E
a. apetnaba	0	0	-1	0.67	0.58
b. aptnaba	-1	-1	0	0.33	0.42

Tableau 5: Variable-Lexical language – known prefixes and stems; /vraba/ exceptionally triggers prefix vowel deletion

Fit with the training data is not as close as with the other languages due to pervasive exceptionality: 20% of the stems (84 morphemes) must utilize scales to capture their behavior, which conflicts with the prior's pressure to keep the total weights and scales low. The effect of the prior is explored systematically in the next section, but it is worth noting here that a closer fit with the training data for this language is straightforwardly achieved with a weaker prior; for example, setting C = 0.1 yields a deletion rate of 97% for /ape-vraba/.

4 Generalizing from exceptional data

This section examines the model's predictions for nonce data, focusing on how the choice of the prior and the rate of exceptionality in the training data affect generalization.

4.1 Effect of the prior

Recall that the previous section reported on experiments with the prior term C set to 1.0. Here, we vary C and examine its effects on the model's predictions, using the Lexical language as a test case.

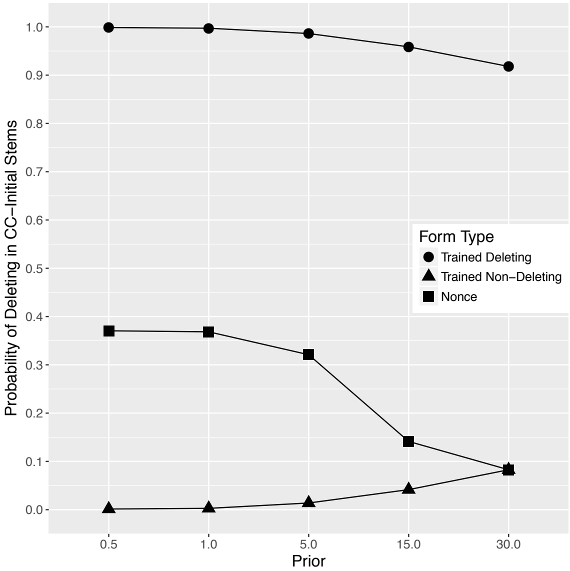

Figure 1: Probability of deletion with CC-stems by C values

Unsurprisingly, the model's fit to the training data decreases as the strength of the prior increases, as there is more pressure to keep all weights and scales close to zero. The model's predictions for novel forms also vary with C, as shown in Figure (1), which plots the probability of deletion for CC-stems by values of C. As C increases, there is a poorer fit to the training data: known forms which should undergo vowel deletion are slightly less likely to, and known forms which should not undergo vowel deletion are slightly more likely to.

The most striking trend is the convergence of nonce form behavior with the behavior of non-deleting forms. In the Lexical language, two prefixes, and thus two-thirds of the data, categorically do not undergo deletion with CC-stems. As discussed in §3.3, these proportions make multiple ways of encoding exceptionality available to the model. When the prior is weak, the model encodes exceptionality in a distributed way, and its predicted deletion rate for novel forms is intermediate between the deleting and non-deleting forms in the training data. When the prior is strong, however, the learner is forced to set more weights to zero, and the non-deleting forms in the learning data are more easily accommodated by the general constraint weights. This leads the learner to designate one of the prefixes as exceptional and to generalize to novel forms on the basis of the non-deleting prefixes. Thus, with a stronger prior, there is more pressure on the learner to over-extend the more general pattern in the data.

This pattern is clear when we examine the learned weights. Table (5) reports the weights learned with C set to 0.5 and 30. With C set low, the learner assigns weight to the exceptionally deleting prefix as well as the non-deleting prefixes. With C set high, the learner only assigns weight to the exceptional prefix, picking it out as exceptional.

	$C = 0.5$			$C = 30$		
	*CCC	Max	Align	*CCC	Max	Align
General Wts	5.3	0.0	4.8	2.4	0.0	0.01
Except. Prefix	0.0	0.0	7.1	0.0	0.0	4.8
Reg. Prefixes	6.0	0.0	0.0	0.0	0.0	0.0
Stems	0.0	0.0	0.0	0.0	0.0	0.0

Table 5: Lexical weights and mean scales, $C = 0.5$ and 30

4.2 Effect of exceptionality

Following Moore-Cantwell and Pater (2016), this section reports the effect of varying the proportion of exceptional forms in the training data on nonce form predictions. To test this, we started with the Categorical language, in which prefix vowels always delete with C-stems but never delete with CC-stems, and then created data sets which increased the percentage of CC-stems that trigger deletion of the prefix vowel by 10% increments, forming a total of 11 data sets (with deletion rates of 0%, 10%, . . . , 90%, 100%). In these simulations, epochs were increased up to 80000 (we found this to be necessary to guarantee convergence for languages with pervasive exceptionality and weaker priors).

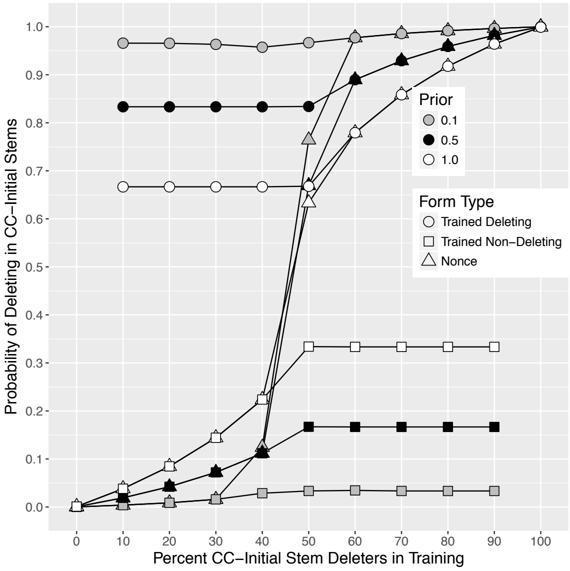

Figure 2: Probability of deletion with CC-stems by percentage of deletion-triggering CC-stems in the training data

Figure (2) plots the probability of deletion as a function of the percentage of triggering CC-stems in the training data. To show how the strength of the prior interacts with the rate of exceptionality in the data, we show curves for three settings of the prior parameter ($C = 0.1, 0.5, 1.0$). The patterns are qualitatively similar for all three settings, with closer fit to the data for weaker priors. As the percentage of triggering CC-stems in the training data increases, the probability of deleting the prefix vowel before any CC-stem increases. For trained stems, the probability of deleting with a non-triggering stem is always much lower than the probability of deleting with a triggering stem, with rates closer to categorical for lower C values. The rate of exceptionality affects learning of both the majority and minority patterns: the more extreme the imbalance, the more poorly the minority pattern is learned and the more categorically the majority pattern is learned.

The behavior of nonce forms mirrors the behavior of non-triggering stems when they form a majority of the training data (0%-30%), and mirros the behavior of triggering stems when they form a majority in the training data (60%-100%), with the probability of deleting in nonce forms rising sharply across those data sets where there is not a clear majority (40%-50%). This indicates that, when there is a clear majority pattern, the model more strongly trends towards using the general weights to capture the majority pattern and the scales to capture the behavior of exceptional forms. When there is no clear majority pattern, the model will trend towards learning general weights which more closely reflect the lexical statistics in the training data, using scales to account for the idiosyncratic behavior of each stem.

Also noteworthy are the non-linear shape and the displacement of the nonce form curves. As Moore-Cantwell and Pater (2016) found for indexed constraints, we show here that the lexically scaled model also predicts nonce form rates that exaggerate the proportions in the training data. For example, when 70% of the stems trigger deletion in the training data, the model with $C = 1.0$ exaggerates this to over 85% in novel forms, and when 30% of stems trigger deletion in training, the model predicts fewer than 15% deletion in novel forms. Our results demonstrate two further influences. First, the exaggeration effect is greater for weaker priors since the curves are overall steeper. Second, the curve is shifted leftward: when 50% of the CC-stems are triggers, the predicted rate of deletion for nonce forms is above 50%, regardless of C. The predicted deletion rate is generally higher than might be expected on the basis of the trained deletion rate in CC-stems alone. The presence of categorically deleting C-stems in the data exerts an independent pressure to weight ALIGN more heavily than MAX, favoring deletion overall. When the $C = 1.0$ model is trained without C-stems (not shown), no skew is predicted: 50% deletion is predicted for nonce forms in the 50% training condition. These results indicate that predictions for nonce forms in one context can be influenced by other processes in the language.

5 Discussion

5.1 Why not L2 regularization?

Existing MaxEnt models of phonology overwhelmingly utilize L2 priors rather than L1 priors (Goldwater and Johnson, 2003; Wilson, 2006; Pater et al., 2012). For our purposes, however, we found that the choice of an L1 prior was crucial. While both L1 and L2 priors penalize higher weights, L1 priors are more effective for learning sparse vectors of weights, with as many zeroes as possible (Yan, 2016). The primary challenge for learning in the lexically scaled MaxEnt framework is to use scales sparingly. This requires a strong pressure to set weights exactly to zero, which an L1 prior provides. L2 priors favor solutions with small weights distributed across many parameters; setting weights to zero is not generally the optimal solution.

$$-\sum_i \log p(i) + \frac{1}{2\sigma^2}(\sum_{\gamma\in\Gamma} w_\gamma^2 + \sum_{\gamma\in\Gamma}\sum_{m\in M} s_{\gamma m}^2) \quad (5)$$

In experiments with an L2 prior, we found there was no weighting of the prior that simultaneously eliminated weights from the scales in languages without exceptionality while satisfactorily accounting for the training data. An example of the weights learned for the Variable language with a weak L2 prior is shown in Table (6). These weights were learned using the standard L-BFGS-B optimizer (Byrd et al., 1995) and the objective function in (5).

	*CCC	MAX	ALIGN
General Weights	3.00	0.00	2.50
Prefix Scales	1.00	0.00	0.80
C-Stem Scales	0.00	0.00	0.08
CC-Stem Scales	0.01	0.01	0.00

Table 6: Variable weights and mean scales, L2 prior, $\sigma^2 = 1$

While predicted probabilities for trained C-stems and CC-stems fit the training data well (Tableau 6), the model makes wide-spread use of scales: all morphemes use scales to some degree even though the Variable language does not require them. Consequently, the weight of ALIGN is not high enough to predict (near) categorical deletion for C-stems. The wide-spread use of scales also prevents the model from generalizing the rate of deletion to novel forms: deletion is predicted to apply more frequently to novel forms. Using a stronger prior (lowering σ), all weights and scales decrease, but weight remains on the scales and the fit with the training data deteriorates. Thus, the L2 prior fails to predict frequency-matching behavior for free variation, predicting skews not only for lexically-conditioned variation (as predicted by the L1 prior) but also for patterns without lexical conditioning.

As discussed earlier, we characterized successful learning in terms of feature selection, using scales only when needed to capture lexical conditioning, and we have shown that the L2 prior does not succeed on this criterion, affecting generalization of categorical, lexicalized, and freely variable processes. However, the extent to which language users encode predictable properties of lexical items is not known, and further behavioral research is needed to understand whether and how generalization of variable and exceptional processes is skewed.

	*CCC	MAX	ALIGN		
/ape-taba/	4.00	0.00	3.38	*O*	*E*
a. apetaba	0	0	-1	0.00	0.03
b. aptaba	0	-1	0	1.00	0.97
/ape-tnaba/	4.01	0.01	3.30	*O*	*E*
a. apetnaba	0	0	-1	0.67	0.66
b. aptnaba	-1	-1	0	0.33	0.34

Tableau 6: Variable language tableau – L2 prior, $\sigma^2 = 1$

5.2 No extra penalty for scales

We found that a general L1 prior was sufficient for keeping the model from overusing scales and generalizing beyond the data. The objective function penalizes weights and scales equally, but scales are more costly when many morphemes require the same scaling. Thus, the pressure against scales follows automatically from their limited utility in the grammar. This contrasts with other approaches to MaxEnt learning of exceptionality, which require either additional priors on some constraints to ensure that the model generalizes (Pater et al., 2012) or distinct phases of learning for general and lexical generalizations (Nazarov, 2018; Shih, 2018).

5.3 Multiple correct solutions

We also found that, when lexical exceptionality is present, there are multiple correct solutions for a given problem with different predictions about generalization. Any model that is tasked with learning both general and exceptional patterns must decide which items in the lexicon are the exceptions and which represent the generalizable pattern. We found that the objective function for our model favors overextending clear majority patterns, but is more ambivalent about what to treat as exceptional given balanced data (§3.3). This ambivalence was modulated by the strength of the prior (§4.1) and the presence of related processes in the data (§4.2).

5.4 Future work

A number of avenues for future work remain. As mentioned in §2, the differences between this approach and lexically indexed constraints (Pater, 2010) remain to be explored, as do differences from alternative models for learning variability and exceptionality (Nazarov, 2018; Shih, 2018). Another natural continuation of this research is to apply the learning paradigm described here to more realistic datasets. Following Pater (2007), Linzen et al. (2013, 489) limit scales to only penalizing exponents of the morphemes they are associated with. This locality condition will be important to incorporate before exploring more complex datasets.

Further investigations of the effect of the prior on generalization are needed. While we investigated the consequences of varying C, our focus was limited to the Lexical language. We found that the data distribution, the strength of the prior, and the existence of related processes in the language already introduce strong pressures on the learner's encoding of exceptionality. In some cases, however, we found the proposed prior did not uniquely favor a single solution. Ultimately, these and other modeling decisions require an understanding of how humans perform under similar learning conditions. Connections with experimental work on how humans generalize variable and exceptional patterns is crucial to defining desirable behavior for any learning model.

6 Conclusion

This paper introduces and tests a method for learning lexically scaled MaxEnt grammars. We show that an L1 prior places strong constraints on the encoding of exceptionality and identify a number of factors that affect the model's performance on training data and generalizations to nonce forms, which can be tested against human behavior.

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Blake Allen and Michael Becker. 2015. Learning alternations from surface forms with sublexical phonology. Unpublished manuscript, University of British Columbia and Stony Brook University. Available as lingbuzz/002503.

Michael Becker and Maria Gouskova. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry*, 47(3):391–425.

Paul Boersma and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32(1):45–86.

Richard H. Byrd, 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.

Andries W. Coetzee and Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31(1):47–89.

Andries W. Coetzee and Joe Pater. 2011. The place of variation in phonological theory. In John A. Goldsmith, Jason Riggle, and Alan C. L. Yu, editors, *The Handbook of Phonological Theory*, pages 401–434. Wiley, 2nd edition.

Mirjam Ernestus and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, pages 5–38.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120.

Bruce Hayes and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23(1):59–104.

Bruce Hayes, Péter Siptár, Kie Zuraw, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85(4):822–863.

Brian Hsu and Karen Jesney. 2016. Scalar positional markedness and faithfulness in Harmonic Grammar. In *Proceedings of the Annual Meeting of the Chicago Linguistic Society*, volume 51, pages 241–255.

Wendell Kimper. 2011. *Competing Triggers: Transparency and Opacity in Vowel Harmony*. Ph.D. thesis, University of Massachusetts Amherst.

Iwona Kraska-Szlenk. 1995. *The phonology of stress in Polish*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03.

Tal Linzen, Sofya Kasyanenko, and Maria Gouskova. 2013. Lexical and phonological variation in Russian prepositions. *Phonology*, 30(3):453–515.

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

Aleksei Nazarov. 2018. Learning within- and between-word variation in probabilistic OT grammars. In Gillian Gallagher, Maria Gouskova, and Sora Yin, editors, *Supplemental Proceedings of the 2017 Annual Meeting on Phonology*, Washington, DC. Linguistic Society of America.

Joe Pater, Karen Jesney, Robert Staubs, and Brian Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 62–71. Association for Computational Linguistics.

Joe Pater. 1996. *Consequences of Constraint Ranking*. Ph.D. thesis, McGill University.

Joe Pater. 2007. The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In Leah Bateman, Michael O'Keefe, Ehren Reilly, and Adam Werle, editors, *University of Massachusetts Occasional Papers in Linguistics 32: Papers in Optimality Theory III*, pages 259–296. GLSA, Amherst, MA.

Joe Pater. 2010. Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker, editor, *Phonological argumentation: Essays on evidence and motivation*, pages 123–154. Equinox, London.

Stephanie S. Shih. 2018. Learning lexical classes from variable phonology. In Yuki Seo and Haruya Ogawa, editors, *Selected Papers from Asian Junior Linguists Conference 2*, pages 1–15. ICUWPL.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 477–485. Association for Computational Linguistics.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30:945–982.

Shi Yan. 2016. L1 norm regularization and sparsity explained for dummies. `https://medium.com/mlreview/l1-norm-regularization-`

and-sparsity-explained-for-dummies-5b0e4be3938a.

Kie Zuraw. 2000. *Patterned Exceptions in Phonology*. Ph.D. thesis, University of California, Los Angeles.

Kie Zuraw. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory*, 28(2):417–472.

Jesse Zymet. 2018. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. Ph.D. thesis, University of California, Los Angeles.