Concurrent hidden structure & grammar learning

Adeline Tan
University of California, Los Angeles
adelinertan@gmail.com

Abstract

The concurrent learning of both unseen structures and grammar is an enduring problem in phonological acquisition. The present study develops a joint model of word-UR-SR triples that incorporates a Maximum Entropy model of SRs conditioned on URs. The learner was presented with word-SR frequencies, and successfully learned the hidden structures and grammars that enabled it to generalize well on test data that were withheld during training. When given an option between acquiring a grammar that supported a rich base analysis and one that didn't, the learner always acquired the grammar that supported rich bases. These results suggest that the preference for acquiring a rich base grammar over a non rich base one is an emergent property of the proposed model.

1 Introduction

In order to fully acquire language, a child has to acquire both the representations and grammar of her language from observed surface forms. Representations include underlying forms, metrical structures, morphological boundaries within words, *etc.* Such representations are absent from the observed data that the child receives, and are thus termed hidden structure. The current study focuses on the learning of hidden structure(s) concurrently with the grammar.

Multiple approaches have been proposed for the concurrent learning of hidden structure and an accompanying constraint-based grammar (Tesar and Smolensky, 2000; Jarosz, 2015; Boersma and Pater, 2016; Rasin and Katzir, 2016; Nelson, 2019). Following Eisenstat (2009), Pater et al. (2012), Staubs and Pater (2016), Nazarov and Pater (2017), and O'Hara (2017), this study incorporates a Maximum Entropy (MaxEnt) grammar (Goldwater and Johnson, 2003) that governs the mapping between hidden structures and surface forms. The current study combines the word-hidden structure mapping with the hidden structure-surface form mapping by utilizing the chain rule of probability theory. This produces a joint word-underlying form-surface form (WORD-UR-SR) model that is compatible with a weighted-constraint grammar of UR-SR mappings. While the model is similar to the ones in Staubs and Pater (2016) and Nazarov and Pater (2017), the current study focuses on learning URs that an analyst would posit, with the learned grammar and lexicon subjected to generalization tasks with wug morphemes.

The model and the learner are introduced in §2 and §3 respectively. We then turn to several schematic languages, the first of which is based on English voicing assimilation (§4). This is followed by a set of six stress languages (§5). Two of the languages within the stress set allow multiple analyses, of which only one analysis supports rich bases (Prince and Smolensky, 2004), thus providing the opportunity to determine whether there is a preference for acquiring the rich base grammar over a non rich base one. The final schematic language is based on English velar softening (§6). §7 concludes.

2 Model

The knowledge whose acquisition will be investigated is knowledge of a particular distribution over WORD-UR-SR triples (*e.g.* <CROC-PL, /kɹɑk+z/, [kɹɑks]>: 99%; <CROC-PL, /kɹɑk+z/, [kɹɑkz]>: .003%; <CROC-PL, /kɹɑk+s/, [kɹɑks]>: .002%; ...). In this paper, WORD[1] represents a sequence of morphemes, and morphemes are represented with uppercase letters.

[1] WORD is also abbreviated WD in this paper.

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 55-64.
Held on-line February 7-9, 2022

The probability of a triple can be rewritten as:

$$Pr(WD, UR, SR) = Pr(SR|WD, UR) * Pr(WD, UR) \quad (1)$$

The first term, $Pr(SR|WD, UR)$, is the probability of an SR for a given WORD-UR pair, and is determined by the traditional phonological constraint grammar. For instance, if Pr([bæŋks]|BANK-PL, /bæŋk+z/) = 0.9, then we should interpret it to mean that the WORD-UR pair <BANK-PL, /bæŋk+z/> is realized as SR [bæŋks] 90% of the time. The model proposed here does not condition the UR-SR mapping on the word. Using the example above, this means that Pr([bæŋks]|BANK$_1$-PL, /bæŋk+z/) = Pr([bæŋks]|BANK$_2$-PL, /bæŋk+z/) = Pr([bæŋks]|BANK$_3$-3SG.PRES, /bæŋk+z/), where BANK$_1$ is the financial institution concept, BANK$_2$ is the river concept, and BANK$_3$ is the concept of turning at an angle. Consequently, $Pr(SR|WD, UR)$ = $Pr(SR|UR)$, and the probability of the WORD-UR-SR triple can be simplified to equation (2):

$$Pr(WD, UR, SR) = Pr(SR|UR) * Pr(WD, UR) \quad (2)$$

Such probabilistic mappings of SRs conditioned on URs (*i.e.* $Pr(SR|UR)$) are computed by virtually all probabilistic constraint-based grammars (*e.g.* probabilistic OT, probabilitic versions of Harmonic Grammar, *etc.*) The current study uses a MaxEnt model, which is a weighted constraint grammar.

Following the traditional phonological MaxEnt model, each UR-SR pair (x, y) is associated with a feature vector, $\vec{v}(x, y)$, which captures the pair's properties. For UR-SR pairs, there are two classes of relevant properties. The first class concerns the form that the SR takes. For example, a feature may be used to track how many pairs of adjacent obstruents of an SR have different voicing values. Such features are known as markedness constraints. The second class of features concerns the mapping between the UR and the SR, and are most commonly used to penalize any changes between the two. These features are conventionally known as faithfulness constraints. Each feature has an associated weight, and the feature weights can be organized into the weight vector $\vec{w}$. The features of the UR-SR pair (x, y) are linearly combined (as in equation (3)[2]) to produce its harmony score, $h(x, y)$. $h(x, y)$ is essentially the weighted sum of the UR-SR pair (x, y)'s features, and is a scalar (rather than a vector).

$$h(x, y) = -(\vec{w} \cdot \vec{v}(x, y)) \quad (3)$$

The MaxEnt model then maps each pair's harmony score to its probability (equation (4)).

$$Pr(SR = y|UR = x) = \frac{e^{h(x,y)}}{Z(x)} \quad (4)$$

Since the traditional phonological MaxEnt grammar is a conditional ("discriminative") model, the partition function $Z(x)$ sums over all UR-SR pairs that share the same UR (equation 5).

$$Z(x) = \sum_{y' \in \mathcal{Y}_x} e^{h(x,y')} \quad (5)$$

In equation (5), $\mathcal{Y}_x$ is the set of all SRs that are compatible with UR x. This has the effect of normalizing the probability of a particular UR-SR mapping among only all other mappings from the same UR.

The second term in equation (2), $Pr(WD, UR)$, is the joint probability of a WORD-UR pair. This implicitly defines a conditional distribution $Pr(UR|WD)$ (equation (6)).

$$Pr(UR = x|WD = w) = \frac{Pr(WD = w, UR = x)}{\sum_{x'} Pr(WD = w, UR = x')} \quad (6)$$

Under this conditional distribution we would expect Pr(/kɹɑk+z/|CROC-PL) to be high, and Pr(/kɹɑk+s/|CROC-PL), Pr(/kɹɑg+z/|CROC-PL), *etc.* to be low. For the morpheme CROC, the learner needs to choose between 2 possible stem-final segments: voiceless /k/ and voiced /g/[3]. For the plural morpheme, the learner needs to choose between voiceless /s/ and voiced /z/. Consequently, there are four potential URs that the learner considers for the word CROC-PL (Table 1). Table 1 also shows the four features for each

[2] A UR-SR pair is active for a phonological constraint when it violates the requirements of that constraint, which in turn reduces the pair's conditional probability. Hence the negative sign in equation (3).

[3] This example is modeled after English voicing assimilation where adjacent obstruents agree in voicing. The surface sequence [ks] could have arisen from any of the following UR sequences {/k+s/, /k+z/, /g+s/, /g+z/}. Hence, I vary only the stem-final segment, but none of the other stem segments.

of the four variants that the learner has to choose among. These features represent the strength of association between a particular morpheme and an aspect (*e.g.* morpheme-final obstruent voicing) of its UR. Within phonology, such features are also known as UR constraints (Zuraw, 2000; Boersma, 2001). Similar to its UR-SR counterpart, there is a feature vector $\vec{u}(w, x)$ for each WORD-UR pair (w, x). Likewise, the UR constraint weights can be organized into a vector $\vec{\theta}$. The harmony score for each WORD-UR pair is computed as per equation (7)[4].

$$g(x, y) = \vec{\theta} \cdot \vec{u}(x, y) \tag{7}$$

The harmony score of a WORD-UR pair is then mapped to its probability (equation 8).

$$Pr(WD = w, UR = x) = \frac{e^{g(w,x)}}{Z} \tag{8}$$

In contrast to the UR-SR model described above, the WORD-UR model is not conditional. The normalization takes place over all WORD-UR pairs (equation (9)).

$$Z = \sum_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}_w} e^{g(w,x)} \tag{9}$$

In equation (9), $\mathcal{W}$ is the set of words, and $\mathcal{X}_w$ is the set of all URs that are compatible with word w. This normalization produces a generative distribution over WORD-UR pairs, which in turn produces the generative distribution over WORD-UR-SR triples of equation (2). This departs from the models in Staubs and Pater (2016) and Nazarov and Pater (2017), which are discriminative models. A generative model is capable of describing differences in the frequencies of various words, in addition to the relationship between words and their realizations, whereas a discriminative model only does the latter.

3 Learning

The model takes in a set of WORD-SR pair frequencies (*e.g.* {<CROC-PL, [kɹɑks]>: 50; <CROC-PL, [kɹɑkz]>: 0; ...}), and learns a probability distribution over WORD-UR-SR triples (*e.g.* <CROC-PL, /kɹɑk+z/, [kɹɑks]>: 99%; <CROC-PL, /kɹɑk+z/, [kɹɑkz]>: .003%; <CROC-PL, /kɹɑk+s/, [kɹɑks]>: .002%; ...). The triple probability defined in Section 2 in fact implicitly defines a distribution over WORD-SR pairs as well. More concretely, the probability of pairs can be computed from the probability of triples via this summation:

$$\begin{aligned} &Pr(WD = w, SR = y) \\ &= \sum_x Pr(WD = w, UR = x, SR = y) \\ &= \sum_x Pr(SR = y | UR = x) * Pr(WD = w, UR = x) \\ &= \sum_x \frac{e^{h(x,y)}}{Z(x)} * \frac{e^{g(w,x)}}{Z} \end{aligned} \tag{10}$$

The likelihood $Pr(WD, SR)$ can be understood as a function of the parameters $\vec{w}$ and $\vec{\theta}$. Experimentation showed that regularization terms did not improve performance in fitting to test data that was withheld from training, so the learner's objective is to seek the values of $\vec{w}$ and $\vec{\theta}$ that maximize this likelihood. In order to assess which values of $\vec{w}$ and $\vec{\theta}$ will be found by the learner, I use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Notice that $Pr(WD, SR)$ is a marginal distribution[5]. The likelihood function of marginal distributions is not guaranteed to be convex, so each EM run finds a local maximum. I take the highest of these local maxima to identify the predicted outcome of learning.

4 English Voicing Assimilation

In English voicing assimilation, adjacent obstruents with different voicing values are resolved with suffixes assimilating their voicing value to that of the stem. The underlying voicing value of stem-final obstruents and suffixes constitute the hidden structures.

4.1 Experimental setup

The first language had the words {CROC-PL, DOG-PL, COW-PL}. Its grammar had the constraints {AGREE(voice), $\text{IDENT}_{\text{stem}}$, $\text{IDENT}_{\text{general}}$}, so $\vec{w}$ was 3-dimensional for this language. In addition, six potential UR variants {/kɹɑk/, /kɹɑg/, /dɑk/, /dɑg/, /-s/, /-z/} were considered, making $\vec{\theta}$ 6-dimensional. These nine dimensions correspond to the first nine rows in Table 3.

What constitutes successful learning? First, we can check whether the UR learned for each morpheme of the training data matches what would

[4] A WORD-UR pair is active for a particular UR constraint when it contains the morpheme, segment, *etc.*, required by that constraint, which in turn increases the pair's probability. Hence the sign difference between equations (3) and (7).

[5] The summation in equation (10) produces marginal probabilities.

WORD	**UR$_{\text{WORD}}$**	(CROC, /kɹɑk/)	(CROC, /kɹɑg/)	(PL, /-s/)	(PL, /-z/)
CROC-PL	/kɹɑk+s/	1	0	1	0
	/kɹɑg+s/	0	1	1	0
	/kɹɑk+z/	1	0	0	1
	/kɹɑg+z/	0	1	0	1

Table 1: UR constraints for the word CROC-PL.

be predicted via traditional phonological analyses. For example, a phonologist would posit that a child learns the URs /kɹɑk/, /dɑg/ and /-z/ for the CROC, DOG and -PL morphemes respectively. Recall that the model produces a distribution over WORD-UR-SR triples. In order to find the probability of the word-sized UR containing /-z/ in the appropriate position, given that the WORD (*i.e.* sequence of morphemes) has -PL in that position, we apply the following equation:

$$p(\text{/-z/}|\text{-PL}) = \frac{\begin{array}{c}[Pr(\text{UR=/dɑk-z/, WD=DOG-PL}) \\ +Pr(\text{UR=/dɑg-z/, WD=DOG-PL}) \\ +Pr(\text{UR=/kɹɑk-z/, WD=CROC-PL}) \\ +Pr(\text{UR=/kɹɑg-z/, WD=CROC-PL}) \\ +Pr(\text{UR=/kaw-z/, WD=COW-PL})]\end{array}}{\begin{array}{c}[Pr(\text{WD=DOG-PL})+Pr(\text{WD=CROC-PL}) \\ +Pr(\text{WD=COW-PL})]\end{array}} \quad (11)$$

The probability of a particular UR for each morpheme is calculated in the same manner for each of the other morpheme-UR pairs.

Second, the model must be able to generalize to unseen data in the way that humans do. Unseen data are WORD-SR pairs that the model wasn't provided with in the training set. For English voicing assimilation, the words in Table 2 provide a good test set for generalizability. Each test word is composed of a new morpheme and an old morpheme. This isolates each of the old morphemes, so that only one old morpheme appears in each word. This allows us to test what the model knows between /-s/ *vs.* /-z/ as the UR of the -PL morpheme, as well as what the model knows about how /-s/ & /-z/ are realized on the surface. The new nouns WUG /wʌg/, HEAK /hik/ & CRA /kɹɑ/ will illuminate what the model had learned about the -PL suffix. I introduce a novel suffix with UR /-d/ (representing some morpheme I will write as -D), to test what the model had learned about the roots DOG and CROC. For example, if $Pr((\text{/hik-z/, [hiks]})|\text{HEAK-PL})$ is high, then we'd know that the model generalizes in the same way that English speakers do, as evidenced by wug tests. The probability of a UR-SR pair for a given word is calculated as per equation (12).

WORD
WUG-PL
HEAK-PL
CRA-PL
DOG-D
CROC-D

Table 2: Test set words for English voicing assimilation.

$$Pr(UR = x, SR = y|WD = w) = \frac{Pr(WD = w, UR = x, SR = y)}{\sum_{x'}\sum_{y'} Pr(WD = w, UR = x', SR = y')} \quad (12)$$

4.2 Results

The training data consisted of 10 logically possible WORD-SR pairs, of which only three were observed. Each of the three observed pairs {(CROC-PL, [kɹɑks]), (DOG-PL, [dɑgz]), (COW-PL, [kawz])} was only observed once. The learner sought the parameter values that maximized the likelihood of the training data. Five settings of the parameters are shown in Table 3. I found these by running the EM algorithm from 20 randomly initialized[6] starting points. The likelihood of the training data for each of the five parameter settings is $0.33 \times 0.33 \times 0.33 = 0.33^3$. These five settings have already hit the maximum likelihood of training data – there isn't another parameter setting that would provide a much better likelihood since these settings have already matched the empirical relative frequencies almost perfectly.

[6]Initial weights for all eight languages in the present study were drawn from a uniform distribution with range=[0.1, 5) for phonological constraints & range=[0, 100) for UR constraints.

[7]In this paper, negative weights were only allowed for UR constraints. Weights for regular phonological constraints were not allowed to be negative. For voicing assimilation, the UR constraint (COW, /kaw/) was excluded from the set of features, since the morpheme COW had only one underlying form under consideration. This resulted in (DOG, /dɑg/) attaining 0 weight, which pushed (DOG, /dɑk/) to a negative weight. Because it is the difference between weights rather than the actual value of the weights that matter, the negative weights do not have any meaningful impact on the results.

	1	2	3	4	5
AGREE(voice)	24.7	40.9	24.4	33.3	26.5
$\text{IDENT}_{\text{stem}}$	15.0	11.9	12.6	29.4	11.6
$\text{IDENT}_{\text{general}}$	11.9	18.5	12.0	18.5	13.6
(DOG, /dɑk/)	-43.2	-11.9	-30.0	-26.4	-11.6
(DOG, /dɑg/)	0.0	0.0	0.0	0.0	0.0
(CROC, /kɹɑk/)	0.0	0.0	0.0	0.0	0.0
(CROC, /kɹɑg/)	-16.3	-23.7	-20.5	-33.1	-13.3
(-PL, /-s/)	0.3	14.4	9.5	20.7	24.6
(-PL, /-z/)	91.0	56.8	76.6	83.0	77.4
Pr(CROC-PL, [kɹɑks])	0.33	0.33	0.33	0.33	0.33
Pr(DOG-PL, [dɑgz])	0.33	0.33	0.33	0.33	0.33
Pr(COW-PL, [kawz])	0.33	0.33	0.33	0.33	0.33
Likelihood of training data	0.33^3	0.33^3	0.33^3	0.33^3	0.33^3
Negative log-likelihood of training data	-3.29585	-3.29585	-3.29585	-3.29584	-3.29586

Table 3: Feature weights[7], probability of observed data, & likelihood of training data from the best five runs (English voicing assimilation).

Recall the first criterion of successful learning: the learner has to learn the very same morpheme-sized URs that human learners are posited by phonologists to learn. Applying the equation in (11), we see that the lexicon learned is indeed the one that matches with traditional phonological analysis (Table 4[8]). A further examination of the UR constraints confirms that the UR constraints associated with expected URs (*i.e.* DOG has underlying root-final /g/, CROC has underlying root-final /k/, the plural morpheme is underlying /-z/) have higher weights than their counterparts (Table 3).

	1	2	3	4	5
p(/dɑg/|DOG)	1.0	1.0	1.0	1.0	1.0
p(/kɹɑk/|CROC)	1.0	1.0	1.0	1.0	1.0
p(/-z/|-PL)	1.0	1.0	1.0	1.0	1.0

Table 4: Lexicon (English voicing assimilation).

To fulfill the second criterion of successful learning, the learned models had to generalize in the same way that English speakers generalize. The generalization task considered 8 UR-SR combinations for each word[9]. For nonce word HEAK /hik/, an English speaker would produce [hiks] for HEAK-PL via the UR /hik-z/. Thus, successful generalization for the word HEAK-PL required assigning high probability to the UR-SR pair (/hik-z/, [hiks]), and low probability to the seven other pairs. Table 5 presents, in the top five rows, Pr((UR, SR)|WD) for the UR-SR pairs that are expected to have high probability for their respective words, given what we know about how English speakers behave on wug tests. The results in Table 5[10] indicate that all five models very successfully generalized in a manner that mimicked speakers, with probabilities close to or at 100%. A look at the learned phonological constraint weights in Table 3 shows why all five parameter settings mirrored speakers so well in the generalization task. The models all learned the two crucial weight-inequalities required for English voicing assimilation: $\text{AGREE(voice)} > \text{IDENT}_{\text{general}}$ as well as $\text{IDENT}_{\text{stem}} > 0$.

For voicing assimilation, the model did well on both the UR-learning of morphemes & wug-test mirroring tasks because both the lexicon & grammar learned by the learner mirrored what English speakers are believed to have learned for voicing assimilation.

5 PAKA stress languages

5.1 Experimental setup

The next 6 languages were generated using similar morphemes and constraints that generated the PAKA World dataset in Tesar et al. (2003). There were two roots: (PA & BA), as well as two suffixes (-KA & -GA). The URs of PA and KA were always unstressed: /pa/ & /-ka/. In contrast, the

[8] All probabilities in Table 4 were very close to or at 100%. The lowest was 99.9991% for p(/dɑg/|DOG) of the fifth parameter setting.

[9] To illustrate the 8 combinations, consider the nonce word WUGS. 2 variations are available via the UR: /-s, -z/, 2 via the SR of the stem-final consonant: [wʌk, wʌg], and 2 via the SR of the suffix consonant: [-s, -z]. This produces $2^3 = 8$ UR-SR combinations.

[10] All probabilities in Table 5 were very close to or at 100%. The lowest was 99.9986% for Pr((/dɑg-d/, [dɑgd])|DOG-D) of the fifth parameter setting.

	1	2	3	4	5
$Pr((/\text{wʌg-z}/,[\text{wʌgz}])\vert\text{WUG-PL})$	1.0	1.0	1.0	1.0	1.0
$Pr((/\text{hik-z}/,[\text{hiks}])\vert\text{HEAK-PL})$	1.0	1.0	1.0	1.0	1.0
$Pr((/\text{kɹɑ-z}/,[\text{kɹɑz}])\vert\text{CRA-PL})$	1.0	1.0	1.0	1.0	1.0
$Pr((/\text{dɑg-d}/,[\text{dɑgd}])\vert\text{DOG-D})$	1.0	1.0	1.0	1.0	1.0
$Pr((/\text{kɹɑk-d}/,[\text{kɹɑkt}])\vert\text{CROC-D})$	1.0	1.0	1.0	1.0	1.0
$Pr((/\text{wʌg-s}/,[\text{wʌgz}])\vert\text{WUG-PL})$	0.0	0.0	0.0	0.0	0.0
$Pr((/\text{wʌg-z}/,[\text{wʌkz}])\vert\text{WUG-PL})$	0.0	0.0	0.0	0.0	0.0
⋮	⋮	⋮	⋮	⋮	⋮

Table 5: Probability of UR-SR pair for a given test set word (English voicing assimilation).

URs of BA & GA could bear stress, so the model considered the potential-URs {/ˈba/, /ba/, /-ˈga/, /-ga/}. Accordingly, the relevant UR constraints for this dataset were: (BA, /ˈba/), (BA, /ba/), (GA, /-ˈga/) & (GA, /-ga/). The four features that went into $Pr(SR|UR)$ were:

- MAINLEFT (ML)
  - Stress the leftmost syllable.
- MAINRIGHT (MR)
  - Stress the rightmost syllable.
- $\text{MAX}_{\text{general}}$-STRESS ($F$)
  - If a syllable is stressed in the UR, retain its stress in the SR.
- MAX_{root}-STRESS (FR)
  - If a root syllable is stressed in the UR, retain its stress in the SR.

These morphemes & constraints produced six logically possible languages[11], which are the six sets of observed SRs shown in Table 6. Languages 3, 4 & 6 were each compatible with only 1 lexicon. Language 5 was compatible with both /-ga/ & /-ˈga/, but compatible with only a single grammar. For Languages 1 & 2, four lexicon-grammar combinations were available for each language.

5.2 Results

For each language, the training data had 12 logically possible WORD-SR pairs[13], of which four pairs were each observed once. The four SRs for each language can be read off column "Observed SRs" of Table 6. As with English voicing assimilation, I did 20 EM runs per language. For all six languages, the learner succeeded in finding multiple parameter settings that hit the maximum likelihood of training data $0.25 \times 0.25 \times 0.25 \times 0.25 = 0.25^4$. For the sake of brevity, only one of these parameter settings is presented for each of the six languages (Table 7).

To test generalizability, two new roots {SO /so/, ZE /ˈze/} and two new suffixes {-FO /-fo/, -VE /-ˈve/} were introduced to form test set words (Table 8). As expected, all parameter settings that attained the maximum likelihood of training data generalized to test words at near 100% probability. A sample of the probabilities of UR-SR pairs for test word BA-FO is shown for one simulation of Language 3, where the combination of trained morpheme BA with an unaccented suffix like /-ka/ produced stress on the first syllable. Likewise, when BA combines with unstressed /-fo/, successful generalization requires a UR-SR pair with [ˈbafo] to have high probability (Table 9).

5.3 Rich base supporting grammars

According to Prince and Smolensky (2004), the role of a constraint-based grammar is to assign an output to each input[15]. In the case of absolute ill-formedness (*e.g.* absence of right-stressed SRs in a left-stressed language), the grammar (*i.e.* the constraint interactions that govern the UR-SR mapping) must ensure that no input ever leads to ill-formed outputs (*e.g.* not even a UR with rightmost

[11] Since MaxEnt generates probabilistic languages, there are technically an infinite number of possible languages. However, I'm restricting the set of languages to only those where there is effectively only one winning SR per UR.

[12] Since the model is MaxEnt rather than non-probabilistic Harmonic Grammar, the difference between the terms on both sides of an inequality need to be sufficiently large in order to generate categorical outcomes. Determining exactly how large a difference is needed for each inequality is difficult. Nevertheless, the test task provides a way to check that the trained weights indeed produce sufficient difference between the two terms of an inequality. If the difference were not sufficiently large, the test task would fail to produce categorical outcomes.

[13] There were three SRs per word – left-stress, right-stress, and no stress at all.

[14] UR constraints for the morphemes PA & -KA that do not have multiple URs under consideration were included in this feature set. Hence negative UR constraint weights do not make an appearance here.

[15] Prince and Smolensky (2004) were writing about Optimality Theory, where the grammar consisted of ranked constraints picking a sole output for each input. Nevertheless, the grammar's role in mapping inputs to outputs still holds for probabilistic constraint-based grammars.

Lg	*Observed SRs*	Description	*Lexicon*	*Required weight inequalities*[12]
1	[ˈpaka, ˈpaga, ˈbaka, ˈbaga]	predictable left-stress	/ba, -ga/ /ˈba, -ga/	ML > MR
			/ba, -ˈga/ /ˈba, -ˈga/	ML > MR + F
2	[paˈka, paˈga, baˈka, baˈga]	predictable right-stress	/ba, -ga/ /ba, -ˈga/	MR > ML
			/ˈba, -ga/ /ˈba, -ˈga/	MR > ML + F + FR
3	[ˈpaka, paˈga, ˈbaka, ˈbaga]	full accentual contrast, default left	/ˈba, -ˈga/	MR + F > ML > MR
4	[paˈka, paˈga, ˈbaka, baˈga]	full accentual contrast, default right	/ˈba, -ˈga/	ML + F + FR > MR > ML + FR
5	[paˈka, paˈga, ˈbaka, ˈbaga]	contrast in roots only, default right	/ˈba, -ga/ /ˈba, -ˈga/	ML + FR > MR > ML
6	[ˈpaka, paˈga, ˈbaka, baˈga]	contrast in suffixes only, default left	/ba, -ˈga/	MR + F > ML > MR

Table 6: PAKA languages and respective logically-possible lexicon-grammar combinations.

	Lg 1	*Lg 2*	*Lg 3*	*Lg 4*	*Lg 5*	*Lg 6*
MAINLEFT (ML)	19.6	0.0	16.9	8.4	0.0	44.7
MAINRIGHT (MR)	0.0	20.0	0.0	23.3	19.4	0.0
$\text{MAX}_{\text{general}}$-STRESS (F)	1.5	1.7	32.4	37.4	21.7	111.4
MAX_{root}-STRESS (FR)	3.1	0.0	7.5	0.0	23.2	2.2
(BA, /ba/)	69.9	47.3	15.5	36.4	27.8	81.1
(BA, /ˈba/)	0.5	77.2	61.4	80.5	66.7	34.2
(GA, /-ga/)	59.6	28.1	11.1	25.8	71.4	59.3
(GA, /-ˈga/)	6.2	65.2	61.2	85.0	35.2	76.8
Likelihood $_{\textit{training data}}$	0.25^4	0.25^4	0.25^4	0.25^4	0.25^4	0.25^4
Negative log-likelihood $_{\textit{training data}}$	-5.545177	-5.545177	-5.545178	-5.545178	-5.545177	-5.545177

Table 7: Feature weights[14] & likelihood of training data from the best runs for each PAKA language.

WORD
SO -FO
SO -GA
BA -FO
ZE -GA
BA -VE

Table 8: Test set words for the six PAKA languages.

WORD	UR-SR pair	Pr(UR, SR|WD)
BA -FO	/ba-fo/, [bafo]	2.7×10^{-12}
	/ba-fo/, [ˈbafo]	4.6×10^{-5}
	/ba-fo/, [baˈfo]	2.7×10^{-11}
	/ˈba-fo/, [bafo]	1.8×10^{-21}
	/ˈba-fo/, [ˈbafo]	9.9995×10
	/ˈba-fo/, [baˈfo]	1.7×10^{-20}

Table 9: Generalization to BA-FO in Language 3 (one run shown).

stress can produce an SR with rightmost stress). Within models that feature probabilistic UR-SR mappings, this translates to the grammar ensuring that no inputs ever map to ill-formed outputs with anything other than a vanishingly small probability. In other words, the grammar should be fail-safe; it should be able to map all URs (even implausible ones like a right-stressed UR in a left-stressed language) to SRs with appropriate probability values. This concept is known as the Richness of the Base (Prince and Smolensky, 2004).

Language 1 & Language 2 are languages with predictable left- and right-stress respectively. Each of these two languages is compatible with two grammars (Table 6). In Language 1, the two possible grammars are Grammar 1 ($ML > MR$) & Grammar 2 ($ML > MR + F$). The lexicon that includes /-ga/ minimally requires Grammar 1, while the lexicon that includes /-ˈga/ minimally requires Grammar 2. Since the weights of phonological constraints could not be negative, Grammar 2 entails Grammar 1. It follows that Grammar 2 is compatible with both /-ga/ & /-ˈga/ while Grammar 1 is compatible with only /-ga/. Grammar 2 is thus a grammar that supports rich bases because it is capable of producing SRs with the right proba-

bilities even with an implausible UR (underlying stressed suffix /-ˈga/ in a left-stressed language where unstressed root /pa/ also exists). In contrast, Grammar 1 is the less restrictive grammar because it requires only $ML > MR$, thus allowing $MR + F > ML$. The gang effect of right-stress (MR) and general faithfulness (F) over left-stress (ML) results in /pa-ˈga/ surfacing with rightmost stress *[pa-ˈga]. For Language 2, this entailment relation also holds amongst its two grammars, with Grammar 4 ($MR > ML + F + FR$) being the rich base supporting grammar, and Grammar 3 ($MR > ML$) the less restrictive one. These two languages thus provide useful test cases on whether there is a preference for a rich base supporting grammar over its less restrictive rival or vice versa.

All 20 runs for Languages 1 & 2 always learned the rich base grammar. Trained weights for an example run of Language 1 is shown in Table 7, where the rich base grammar, ML (19.6) > MR + F (0 + 1.5), is learned. To test this further, I ran 200 more simulations for both languages. All 200 runs for both languages always learned the rich base supporting grammar, sometimes with the lexicon that minimally required the rich base grammar, and sometimes with the lexicon that minimally required the less restrictive one. This indicated a strong preference for learning the rich base grammar over its less restrictive counterpart.

The preference for acquiring the rich base grammar is an emergent property of the model. EM finds the local maximum by hill-climbing from a randomly initialized point within the solution space. The solution space is the likelihood function of the marginal distribution (equation (10)) of the model defined in §2. Hill-climbing (*i.e.* gradient ascent) is guided by the gradients of the solution space at the current point. The preference for converging at maxima corresponding to the rich base grammar indicates the following: within the solution space, there are more points with gradients pointing towards maxima corresponding to the rich base grammar and fewer points with gradients pointing towards maxima corresponding to the less restrictive grammar. Since the solution space is a property of the model (rather than that of a particular learner), models with similar architecture (*e.g.* Staubs and Pater (2016); Nazarov and Pater (2017)) are likely to also favor the acquisition of rich base grammars.

6 Velar Softening

In English velar softening, /k/ → [s] before a high front vowel when a morpheme-boundary intervenes (*e.g.* *electri*[k]∼*electri*[s]-*ity*). Velar softening is an instance of the derived environment effect (DEE) because its triggering environment requires the presence of a morpheme boundary. DEEs are a puzzle because both the alternation and the segmentation into morphemes must be acquired simultaneously. In an additional wrinkle, DEEs are often only triggered by specific morphemes. For example, the *-ity* morpheme triggers velar softening, but the *-ish* morpheme does not. Velar softening thus has three sources of hidden structure – presence of a morpheme boundary, whether a particular suffix is exceptional in triggering velar softening, and the usual UR-segment-learning (/k/ or /s/) that we've already seen in the preceding test cases. I use the *-symbol to indicate the exception tagged UR variant.

6.1 Experimental setup

There were eight observed words {ELECTRIC, ELECTRICITY, ELECTRICISH, KITTY, SECURE, SECURITY, SMALL, SMALLISH}. The three sources of hidden structure were combined to produce URs for these eight words. The URs for ELECTRICITY are shown with their relevant UR constraint features (Table 10). The UR /electrik-*ity/ contained the morphemes ELECTRIC[16] & -ITY, so it was active for those two features. These features represent a new class of UR constraint, namely those that indicate the presence of certain morphemes. Such features were not required in the preceding cases as the morpheme boundary was not in question. The UR /electrik-*ity/ had underlying /k/ for morpheme ELECTRIC, and the exception-tagged version of the -ITY suffix, so it was also active for features (ELECTRIC, k)[17] & (-ITY, -*ity) respectively. These UR constrains are the same kind that we've seen before.

Three phonological constraints controlled the UR-SR mapping – a general markedness constraint against [kɪ] sequences (M), an exception-tagged version that was active only when an exception-tagged morpheme was part of the [kɪ] sequence (ME), and a general IDENT constraint (F_{gen}).

[16] Abbreviated as EL...C in Table 10.
[17] Abbreviated as (EL...C, k) in Table 10.

UR$_{\text{ELECTRICITY}}$	EL...C	EL...CITY	-ITY	(EL...C, k)	(EL...C, s)	(EL...CITY, k)	(EL...CITY, s)	(-ITY, -*ity)	(-ITY, -ity)
/electrik-*ity/	1	0	1	1	0	0	0	1	0
/electrik-ity/	1	0	1	1	0	0	0	0	1
/electris-*ity/	1	0	1	0	1	0	0	1	0
/electris-ity/	1	0	1	0	1	0	0	0	1
/electrikity/	0	1	0	0	0	1	0	0	0
/electrisity/	0	1	0	0	0	0	1	0	0

Table 10: URs under consideration for the word ELECTRICITY shown with their UR constraint features (English velar softening).

6.2 Results

The training data had 12 logically possible WORD-SR pairs, of which eight were each observed once. Of 125 EM runs, multiple parameter settings were found to have reached the maximum likelihood of training data $0.125^8 = 5.9605 \times 10^{-5}$. 90.9% of these parameter settings learned the very same hidden structures that matched the standard phonological analysis of velar softening. This included learning that the -ITY morpheme was exception-tagged but that -ISH wasn't. All of these hidden features were learned at probabilities[18] greater than 97%, with the lowest going to the probability of a morpheme boundary in SECURITY at 97.3%.

To test whether the learned grammars generalized in a way that mimicked human learners, the following three morphemes were introduced: a new root with a morpheme-final-/k/ (CLEMIC /ˈklɛmɪk/), an exception-tagged suffix *-ISM /*-ɪzm/ and a non-exception-tagged suffix -Y /-i/ to create the test words in Table 11. For each

WORD	Expected UR-SR
ELECTRIC-ISM	electri/k/-*ism, electri[s]ism
ELECTRIC-Y	electri/k/-y, electri[k]y
CLEMIC-ITY	clemi/k/-*ity, clemi[s]ity
CLEMIC-ISH	clemi/k/-ish, clemi[k]ish

Table 11: Test set words for English velar softening & UR-SR pairs expected to be learned by human learners.

test word, its expected UR-SR pair arose from what traditional phonological analysis would have a child positing as its UR and SR. For instance, the word ELECTRIC-ISM would be posited to have morpheme-final /k/ as opposed to /s/ for ELECTRIC, with that /k/ surfacing as [s]. The expected UR-SR pair for each word is shown in column "Expected UR-SR" of Table 11. The same 90.9% of parameter settings that hit the maximum likelihood of training data generalized well to the test set with probabilities of the expected UR-SR pair for each word approaching 100%[19]. Examination of the weights learned for the parameter settings that successfully generalized confirmed that they had each learned the grammar necessary for velar softening. That is, the alternation applied when the suffix was exception-tagged ($ME + M > F_{gen}$), but did not take place when the suffix wasn't exception-tagged ($F_{gen} > M$).

What would this high-but-not-100% rate of acquisition of the velar softening grammar mean for human learners? Perhaps 10% of people fail to learn the velar softening grammar, and instead rely on memorized forms for existing words. These people are predicted to not apply velar softening in a wug test. Interestingly, in a wug test with nonce stems and the *-ity* suffix, Pierrehumbert (2006) found that 2 in 10 subjects did not have productive velar softening.

7 Conclusion

The present study produced a domain-general model that concurrently learned both hidden structure and a weighted-constraint grammar. The model was trained on eight languages, and generalized well to test data on all of them. Two languages in particular presented a choice between acquiring a grammar that supported rich bases versus one that didn't. This study found a strong preference for acquiring the rich base grammar, which I argued was an emergent property of the model. The present study thus presented a way in which a rich base grammar may be acquired when URs are not known in advance.

Acknowledgments

I would like to thank Tim Hunter and Kie Zuraw for stimulating discussion which made this paper better. I also thank the three anonymous reviewers for helpful feedback.

[18] These probability values were calculated using the same method shown in eq (11).

[19] The lowest probability was $Pr(/\text{klɛmɪk-ɪʃ}/, [\text{klɛmɪkɪʃ}]|\text{CLEMIC-ISH}) = 99.983\%$.

References

Paul Boersma. 2001. Phonology-semantics interaction in OT, and its acquisition. In Robert Kirchner, Wolf Wikeley, and Joe Pater, editors, *Papers in Experimental and Theoretical Linguistics*, volume 6, pages 24–35. University of Alberta, Edmonton.

Paul Boersma and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for harmonic grammar. In John J. McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox Press, London.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B (Methodological), 39(1):1–38.

Sarah Eisenstat. 2009. Learning underlying forms with maxent. Master's thesis, Brown University.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Östen Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–20.

Gaja Jarosz. 2015. Expectation driven learning of phonology, ms. University of Massachusetts, Amherst.

Aleksei Nazarov and Joe Pater. 2017. Learning opacity in stratal maximum entropy grammar. *Phonology*, 34:299–324.

Max Nelson. 2019. Segmentation and UR acquisition with UR constraints. *Proceedings of the Society for Computation in Linguistics*, 2:60–68.

Charlie O'Hara. 2017. How abstract is more abstract? learning abstract underlying representations. *Phonology*, 34:325–345.

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

Janet Pierrehumbert. 2006. The statistical basis of an unnatural alternation. In Louis Goldstein, D. H. Whalen, and Catherine T. Best, editors, *Laboratory Phonology VIII, Varieties of Phonological Competence*, pages 81–107. Mouton de Gruyter, Berlin.

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

Ezer Rasin and Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry*, 47(2):235–282.

Robert Staubs and Joe Pater. 2016. Learning serial constraint-based grammars. In John McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox Press.

Bruce Tesar, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani, and Alan Prince. 2003. Surgery in language learning. In G. Garding and M. Tsujimura, editors, *WCCFL 22 Proceedings*, pages 477–90. Cascadilla Press, Somerville, MA.

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

Kie Zuraw. 2000. *Patterned exceptions in phonology*. Ph.D. thesis, UCLA.