How abstract is more abstract?

Abstract underlying representations can be learned through emergent feature economy

Charlie O’Hara

March 1, 2016

Abstract

This paper presents a Maximum Entropy Learner of Grammars and Lexicons (MaxLex) and demonstrates that this type of system has an emergent preference for minimally abstract underlying representations. In an effort to keep faithfulness constraints weighted low, the learner attempts to fill gaps in the lexical distribution of segments and to make the underlying segment inventory more feature economic. Even when the learner only has access to individual forms, properties of the entire system are implicitly available through the relative weighting of constraints. These properties lead to a preference for some abstract underlying representations over others, mitigating the computational difficulty of searching a large set of abstract forms. These results are illustrated using learning simulations based on the [i]-[∅] alternation in Klamath verbs. This pattern cannot be represented or learned using concrete underlying representation, but MaxLex successfully learns both the phonotactic patterns and minimally-abstract underlying representations.

Regardless of theoretical assumptions, at some point morphemes must be linked to phonological forms. But how distinct can these lexically stored phonological forms be from the forms that actually surface?

A strong hypothesis would claim that for each morpheme there is a single surface form that serves as the underlying representation (UR). If this were true, the surface form of the morpheme in all contexts would be entirely predictable from one surface form. However this is much too strong; some morphemes have alternants that are not predictable in this way. Consider [ˈfɒtəɡræf]-[ˈfɒtəɡrəfi] (photograph-photography) among other forms in English; the vowel place features when stressed are totally unpredictable from the reduced unstressed form (Schane, 1974). Similar patterns are found in Palauan (Schane, 1974; Flora, 1974) and elsewhere. However, both of these patterns can be modeled as long as lexically stored forms are allowed to contain material from multiple surface forms, i.e. /ˈfɒtəɡræf/. This extension from the strong hypothesis has been widely accepted, leading some to draw a line between concrete and abstract URs (Definitions adapted from Kenstowicz & Kisseberth 1979; Baković 2009; Bowers 2015).
1. A *concrete* UR is one such that each feature in the UR appears in at least one of its surface exponents.

2. An *abstract* UR is one that is at least one feature or component in the UR never appears in any of its surface exponents; in other words any non-concrete UR is an abstract UR.

When does a single concrete UR fail to represent an alternation? An alternation cannot be represented by any single concrete UR if all potential concrete URs map to other paradigms in the language. Famous examples of these come from the literature on exceptionality—and while I surrender that some forms truly are exceptional—others cannot only best be analyzed using abstract stored forms, but those abstract forms are easily learnable. As the prototypical example, this paper uses an alternation between [i]-[∅] in Klamath. Author, in press shows that surface distributional patterns suggest that this alternation is underlingly represented abstractly by /e/. This alternation is reminiscent of the better known Slavic yer pattern (Jarosz, 2005; Gouskova & Becker, 2013), but with some additional evidence for an abstract UR. In Klamath, some [i] are nonalternating and appear in all members of their morpheme’s paradigm, but other [i] alternate with nothing ([∅]) when paired with a certain phonological class of suffixes. The alternating [i] is unlikely to be represented underlingly with /i/, because then the nonalternating [i] could not be. Thus, the concrete representation for the segment, /i/, fails to model the alternation. The only remaining concrete option is that the alternating [i] is epenthesised when it appears, which also fails in Klamath.

Some have argued (Kiparsky 1968; Albright 2002; Allen & Becker 2015) that languages should not make use of abstract URs, citing the difficulty for the learner. Each morpheme has a clearly limited number of possible concrete URs, correlated to the number of surface alternants and the length of those. If URs do not need to be concrete, what prevents this set of possibilities from becoming exponentially larger or even infinite? Given that the grammar is output-driven (Tesar, 2014), structure can be applied to this set of URs, searching those with minimum faithfulness violations first. However, since this pattern involves deletion, the set of URs is still very large. What could tell us that the Klamath [i]-[∅] alternation is represented underlingly by /e/, which has a gap in its surface distribution in the position where the alternation is found, and not /i/, /ɪ/ or some diacritic /i_2/ which have no visible effect on the surface?

The difference for the analyst is relatively obvious. Seeing a gap in the distribution of one phoneme, and an alternation that appears only in that gap, the analyst tries to plug that gap; rather than proposing an arbitrary new phoneme for the alternation. This notion is deeply tied into the idea of feature economy, first introduced in (de Groot, 1931, 121): “there is a tendency to employ certain accompanying phoneme properties more than once”\(^1\) If the analyst was to introduce a new phoneme say /i/ to Klamath to explain the variation, they must add in a new contrasting feature, used just once.

While it would be abstract as a representation for the [i]-[∅] alternation, /e/ surfaces faithfully elsewhere in the language. I call /e/ *restrictedly abstract* (3), because it is abstract only in restricted contexts. Since the set of positions within a UR where /e/ would cause the UR to be abstract is a more restricted set of positions than [i], When comparing two URs that differ only on /e/ and /i/, the one with /e/ is more restrictedly abstract than /i/.

\(^1\)Translation from Clements (2003).
Consider two abstract URs for some alternating morpheme, $x=/x_1...x_n/$ and $y=/y_1...y_n/$. Since both are abstract, there must exist some segments in $x$, say $x_i$, so that $x_i$ never appears in a surface exponent of the morpheme; and the same is true for $y_j$ of $y$. $x$ is more *restrictedly abstract* than $y$ for that morpheme, if $x_i$ has a wider surface distribution than the $y_i$ in the entire system of the language.

a. If $/?e:w/\text{ and }/?e:wi/$ are both abstract morphemes being considered for a [?e:w]-[?e:wi]; we look at the distribution of [e] and [i].

b. If the contexts where [e] appear are a superset of those where [i] appears, [?e:we] is more restrictedly abstract than [?e:wi].

Feature economic notions have a long pedigree in phonological analysis (see Clements (2003) §1.5 for a historical overview and citations; esp. Hockett (1955); Martinet (1968)). However since Optimality Theory and other related constraint based grammars tend to have no restrictions on the input or morpheme structure constraints, there is no explicit place for feature economy in the grammar. Yet, grammar is not the only way to obtain language universals and tendencies—recent work has begun to show the power of learnability to restrict typology, through emergent properties (Alderete 2008; Heinz 2010; Staubs 2014; Pater & Staubs 2013; Hughto et al. 2015; Staubs et al. 2016; Pater 2016; Stanton in press, a.o.).

This paper claims that the preference to use a UR that is more restrictedly abstract emerges naturally out of a learner like that typically used in the MaxEnt learning literature (Goldwater & Johnson (2003); Wilson (2006); Hayes & Wilson (2008); Jäger & Rosenbach (2006); Jäger (2007), a.m.o). This emergent bias prefers the choice of URs that minimise gaps in the lexical distribution of segments, without any direct pressures to do so. While it is a largely unsettled question where the line should be drawn between a single UR and allo-morphy, this paper shows that abstract URs can be learnable, allowing for more alternations to be explained with single URs.

Section 1 will introduce the data from Klamath, and show analytically that a concrete UR will fail, and thus an abstract UR may be useful. In section 2, the MaxLex learning model will be presented. In section 3, I will show the results of phonotactic learning and the weighting conditions necessary regardless of the selected UR. In section 4, the simulation results will show that a MaxLex learner learns abstract URs and learns the most restrictedly abstract UR. Section 5 offers explanation as to why the learner prefers certain types of abstract URs to others. Finally in section 6, I conclude and point to further questions.

1 Abstractness in Klamath

In this section, I will sketch out the data from Klamath (Penutian; Southern Oregon) that motivates the learning of some abstract URs. All Klamath data comes from Barker (1964, 1963). For more details on this analysis; see Author (in press). Klamath has a four vowel place contrast on the surface, [i e a u], and contrastive length for each vowel (Barker, 1964,

---

2Pater & Staubs (2013) most closely ties into this work; showing that output-visible feature economy and contrast emerge from iterated learning. In contrast this paper focuses on the analytical predictions of feature economy on the input.
However in noninitial syllables of verbs, short [e] never surfaces. In exactly this context, a set of verb stems exhibit an alternation between [i] and [∅] (4).

(4) Verb stems exhibiting [∅] ~ [i] alternation
With suffix /-a/ ‘indicative’  With suffix /-tkʰ/ ‘having been …ed’
(a) [ʔɛ:w-a] ‘is deep’  [ʔɛ:wi-tkʰ] ‘deep’
(b) [q’ɛ:l’-a] ‘acts silly’  [q’ɛ:l’-i-tkʰ] ‘one acting silly’
(c) [nupu:s-a] ‘burns a little’  [nupusi-tkʰ] ‘charred’
(d) [qu’i’-a] ‘feels passionate’  [qu’i’-tkʰ] ‘passionate’

Thus two potential concrete URs are available, for a morpheme like [ʔɛ:wə]-[ʔɛ:witkʰ]: /ʔɛ:w/ where the [i] is epenthesised, and /ʔɛ:wi/ where the /i/ is deleted. However, neither of these analyses are possible given the rest of the language’s phonology. First, we see that the epenthetic vowel in Klamath is [a] not [i], (5). Thus, if /ʔɛ:w/ were the UR, /ʔɛ:w-tkʰ/ would surface as [ʔɛ:watkʰ] incorrectly.

(5) Verb stems showing /a/ as the default epenthetic vowel
(a) [sk’ɑ:w-a] ‘is cold’  [sk’ɑw-ɑ-tkʰ] ‘cold’  [sk’ɑw-tkɪ] ‘turns cold’
(b) [kuw-ɑ] ‘swells up’  [kuw-ɑ-tkʰ] ‘swollen’  [kuw-y’asq] ‘venereal disease’

Second, many other verb stems have nonalternating [i], as in (6). If /ʔɛ:wi/ were the UR, /ʔɛ:wi-a/ would incorrectly surface as [ʔɛ:wi]. Therefore neither of these concrete URs can model the alternation.

(6) Verb stems with non-alternating stem-final /i/
(a) [stupwɪ] ‘has first menstruation’  [stupwɪ-tkʰ] ‘one who reached womanhood’
(b) [sla:mq’-ɪ] ‘is a widower’  [sla:mq’-i-tkʰ] ‘widower’

The [i]-[∅] alternation cannot be represented concretely, so a learner or analyst might attempt to represent it abstractly. Since [e] never appears in this context, but appears elsewhere, it is a particularly good potential UR, whereas a segment like /i/, which never surfaces in Klamath, seems less good. This is because /e/ is more RESTRICTEDLY ABSTRACT than /i/ as defined above in (3). [e] has a wider distribution than [i], surfacing in a superset of its positions. Since [i] never surfaces, its distribution is the empty set, and thus any segment that ever surfaces appears in a superset of its positions.

Importantly, /i/, and Id(ATR) are merely stand-ins for any never-surface-apparent distinction between the UR and a concrete UR. If /e/ is preferred by the learner to /i/, it would also be preferred to /i’/ or /y/, or even a more diacritic approach. A indexed constraint account, (Benven 1997; Alderete 1999; Itô & Mester 1999; Pater 2000, 2005; Coetzee 2009; Becker 2009; Coetzee & Pater 2011; Gouskova & Becker 2013, a.m.o.) could be such an example, as somehow the relevant morpheme must be marked exceptional.

In the following sections we will see that the restrictedly abstract URs emerge as a naturally preferred UR for the learner in cases like Klamath. This emergent property shows the learner replicating the analyst’s intuitions.
2 MaxLex Learner

The Maximum Entropy Learner of Grammars and Lexicons (MaxLex) used in this paper is built up of a number of aspects used by previous learners. MaxLex uses a MaxEnt Harmonic Grammar as its model of constraint interaction (Goldwater & Johnson (2003); Wilson (2006); Hayes & Wilson (2008); Jäger & Rosenbach (2006); Jäger (2007), a.m.o.). Following previous literature (Hayes, 2004; Jarosz, 2006; Tessier, 2007; Jesney & Tessier, 2011; Tesar, 2014; Alderete et al., 2005; Merchant, 2008; Magri, 2012), the learner has two stages, first a phonotactic level, where the learner is unaware of morphological components of words, and simply tries to find a mapping from surface forms to themselves; and then a morphologically aware stage, where the learner attempts to learn alternations and (in our case) the underlying representations of morphemes. Differentiating this learner from many other MaxEnt learners of underlying forms, which use UR constraints (Eisenstat, 2009; Pater et al., 2012; Staubs & Pater, in press), MaxLex learns a probability distribution across a set of possible URs as the Maximum Likelihood Learner of Lexicons and Grammars (MLG, Jarosz 2006) does in an OT framework.

The learning algorithm implements a batch learner that minimises an objective function. This objective function is built of two parts. The negative log likelihood of the observed data from the phonotactic stage appears in the objective function in (7); predictably not having any term representing the probability of input forms, as the learner is not morphologically aware at this point. This is no different than the objective function typically used in the MaxEnt learning literature cited above.

\[(7)\quad \text{Objective function for first stage of MaxLex}^3\]

\[O_{Ptac}(w) = -ln\left(\prod_{o_i \in \text{SurfForms}} \left( \frac{\prod_{z \in \text{Cand}(o_i)} e^{H(o_i/-[o_i])}}{\sum_{z \in \text{Cand}(o_i)} e^{H(o_i/-[z])}} \right) \right) + \sum_{w_i \in w} \frac{(w_i - c_i)^2}{\sigma_i^2}\]

The objective function on the second level once morphological awareness has kicked in, shown in (8), defines the negative log likelihood of observed data differently. The learner now tries to find the value of the constraint weights \((w)\) and the probability distribution of URs \((\pi)\) that maximises the likelihood of the observed data. The critical changes here are that now instead of just looking at the surface forms of the data, \(o_i\), the learner is aware of a set of morpheme tags in the word; \(\{\mu_{i1}...\mu_{in}\}\). The Harmony and Candidate functions are now evaluated over concatenated strings of possible URs for morphemes, \((u_{i1}...u_{in})\). Since on this stage the learner must find the likelihood of the data given any possible UR, the learner sums up the likelihood of the output form given some UR \((u_{ij})\) for a morpheme \((\mu_{ij})\) and multiplies this by the probability that \(u_{ij}\) is the underlying form for \(\mu_{ij}\) \((P(u_{ij}|\mu_{ij}))\). For this Klamath

\(^3\text{Here } H \text{ represents the harmony score of an input-output candidate, given weights } w. \text{ Cand}(o_i) \text{ is the set of candidates given input } o_i. \text{ } c_i \text{ returns 100 if the } i\text{th constraint is a markedness constraint, and 0 if it is faithfulness. } \sigma_i \text{ is a plasticity constant that is set separately for markedness and faithfulness constraints.}\)
data, each word is only composed of two morphemes, so only $\mu_1$ and $\mu_2$ are important for this paper and the simulations within. The objective function for the morphologically aware stage is thus primarily the same as (7), except the likelihood of each input-output mapping is multiplied by the probability of each UR used in the input.

$\text{Objective function for second stage of MaxLex}$

\[ O_{\text{Lex}}(w, \pi) = -\ln \left( \prod_{\{\mu_1, \ldots, \mu_n\}, o_i \in \text{Data}} \left[ \sum_{u_{i1} \in \text{UR}(\mu_{i1})} \sum_{u_{in} \in \text{UR}(\mu_{in})} P(u_{in}|\mu_{in}) \left( \frac{e^{H(u_{i1} \ldots u_{in} / -[o_i])}}{\sum_{z \in \text{Cand}(u_{i1} \ldots u_{in})} e^{H(u_{i1} \ldots u_{in} / z)}} \right) \right] \right) \]

\[ + \sum_{w_i \in w} \frac{(w_i - c_i)^2}{\sigma_i^2} \quad \text{Negative Log Likelihood} \]

For the simulations in this paper, in order to simplify the data and set of constraints, a toy version of Klamath, Klamath', was used. Klamath' differs from Klamath by using verbal suffix /-ta/, which resembles the combination of the /-a/ suffix seen above and a locative morpheme /-t/ ‘on’\(^4\), will be treated as fully productive. This suffix appears with the [∅] form of alternating verbs (9), and does not cause epenthesis in consonant final stems.

\[ (\text{a}) \quad [n-t\text{jewta}] \ ‘splinters on’ \quad [w-t\text{f}erwit\text{k}] \ ‘pl. broken’ \]
\[ (\text{b}) \quad [nqutta] \ ‘scorches onto’ \quad [nqut\text{t}at\text{k}] \ ‘scorched’ \]

In Klamath, /e/ surfaces faithfully when it is long or when it is in a noun stem. Since more vowel place contrasts are maintained in these positions, they are privileged positions in Klamath. Therefore positional faithfulness constraints (Beckman, 1998; Smith, 1997) to prevent a repair of /e/ in these positions must be active. Long [e:] shows an alternation with [∅] in the same context, showing that mid vowels must be marked in those positions, and that deletion does occur. The toy language is also missing a glottalisation alternation that pairs with the [i]-[∅] alternation in many stems that offers evidence against an epenthesis account. Note all of these differences should make learning that /e/ is the UR in the natural Klamath easier than learning that /e/ is the UR in the toy Klamath’. It follows that if a learner can still learn that /e/ is better than /i/ in the toy language, it should be able to learn the same in the natural language.

3 Stage 1: Phonotactic Learning of Klamath

In stage 1, the learner is presented with the Klamath’ data as shown in (10). One set of forms shows the troublesome alternation that is the focus of this inquiry (10-(a)). The other two are necessary to show the learner that the concrete URs cannot serve as the UR for the alternation; (10-(b)) shows that /i/-final stems fail, and (10-(c)) show that consonant final stems fail.

\[ 4\text{In Klamath this morpheme has a number of alternations that will be ignored for this paper; see Barker 1963, p. 285 for more details.} \]
The Candidate function is implemented by taking in the input and returning all permutations of the output with any violations of the faithfulness constraints involved in the simulation. The only restriction is that epenthesis is limited to occurring between consonants at morpheme boundaries; as to prevent infinite epenthesis and allow us to get by with just one simple cluster constraint. This restricts GEN to only consider the candidates that differ on the constraints currently being considered. The simulation then evaluates the harmony of each input-output candidate by automatically incurring violations of the constraints in question, and multiplies the violations by the weight of the constraint.

The simulations uses a Sequential Least Squares Programming (SLSQP; Kraft 1988) an optimisation algorithm, implemented in SciPy, in order to find the constraint weights (and in stage 2 the lexical probabilities) that lead to the minimal objective function.

The constraints used in the simulations are shown in the first column of the table in (11). The second column shows the initial weights of each constraint, which are all set uniformly at 50. In the third column, the weights output by the learner after convergence at this stage are shown. In the final column, whether a constraint is considered a markedness or faithfulness constraint is marked, in order to show what the constraint was biased to, 100 or 0 respectively.

<table>
<thead>
<tr>
<th>Constraint Weights Learned by Phonotactic Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constraint</td>
</tr>
<tr>
<td>ID(Hi)</td>
</tr>
<tr>
<td>ID(Hi)/σ₁</td>
</tr>
<tr>
<td>MAX-V</td>
</tr>
<tr>
<td>MAX-V/σ₁</td>
</tr>
<tr>
<td>*Mid</td>
</tr>
<tr>
<td>ID(ATR)</td>
</tr>
<tr>
<td>ID(ATR)/σ₁</td>
</tr>
<tr>
<td>Dep-V</td>
</tr>
<tr>
<td>*₁</td>
</tr>
<tr>
<td>PhTac</td>
</tr>
</tbody>
</table>

Most constraints act as would be expected, with a few worth calling attention to. For the purposes of this paper, positions in regards to positional faithfulness constraints are defined upon the input. In order to model positional asymmetries in deletion patterns we must include MAX-V/σ₁ which fails to ever be violated if positions are defined on the output (See Jesney (2015) for more arguments on this front.) The constraint PhTac is a cover constraint here, used to represent overall Klamath phonotactics; it assigns a violation mark to word final clusters of [Ctkʰ] (unless the C is [n].).

3.1 Necessary Weightings in HG for Klamath’

There are two types of weighting conditions involved in modeling the [ı]-[∅] alternation; those that describe the phonotactics and therefore must be true regardless of the choice of UR;
and those that describe the function between inputs and outputs, and allow some UR to
surface with the alternation. The weightings learned on the first non-morphologically aware
level, represent most of the non-UR specific weightings.

The relevant properties that are true of the phonotactics in Klamath’ are:

• [e] does not surface in noninitial syllables
• [i] surfaces nowhere
• [Ctk] clusters do not appear word-finally

First, for [e] to not surface in noninitial syllables, at least one of \( \text{Id(hi)} \) or \( \text{Max-V} \) must be weighted below \( \ast \text{Mid} \) in order to drive some repair of noninitial [e] vowels. In fact, using the weights learned by the phonotactic stage (as in (11)), it is true for both constraints (12). Since [nu.pusi.ta] and [nu.pusi.ta] are so close in weightings, in MaxEnt they each share around half the output probability. Crucially the candidate [nu.pu.se.ta], with the [e] appearing unlicensed in a nonprivileged position, gets near zero, so it almost never surfaces and is marked with a \( \times \).

(12) \( \ast \text{Mid outweights Id(high) and/or Max-V} \)

<table>
<thead>
<tr>
<th>/mupuseteta/</th>
<th>( \ast \text{Mid} ) w = 74.43</th>
<th>( \text{Id(HIGH)} ) w = 40.12</th>
<th>( \text{Max-V} ) w = 40.11</th>
<th>H</th>
<th>( \sim \text{PROB} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \times ) a. nu.pu.se.ta</td>
<td>-1</td>
<td></td>
<td></td>
<td>-74.43</td>
<td>6e-16</td>
</tr>
<tr>
<td>b. nu.pu:si.ta</td>
<td></td>
<td></td>
<td></td>
<td>-40.12</td>
<td>.5</td>
</tr>
<tr>
<td>c. nu.pus.ta</td>
<td>-1</td>
<td></td>
<td>-40.11</td>
<td>.5</td>
<td></td>
</tr>
</tbody>
</table>

[e] surfaces faithfully in initial syllables. Since [e] does not raise in initial syllables, the sum of the weights of \( \text{Id(HIGH)} \) and \( \text{Id(HIGH)}/\sigma_1 \) must be more than the weight of \( \ast \text{Mid} \). To prevent it from deleting, Max-V and Max-V/\( \sigma_1 \) must collectively outweigh \( \ast \text{Mid} \). Both of these results are shown to be learned in (13).

(13) \( \text{Id(HIGH)} + \text{Id(HIGH)}/\sigma_1 \) and \( \text{Max-V + Max-V}/\sigma_1 \) outweigh \( \ast \text{Mid} \).

<table>
<thead>
<tr>
<th>/sn’ewlita/</th>
<th>( \ast \text{Mid} ) w = 74.43</th>
<th>( \text{Id(HIGH)} ) w = 40.12</th>
<th>( \text{Id(HIGH)}/\sigma_1 ) w = 40.12</th>
<th>( \text{Max-V} ) w = 40.11</th>
<th>( \text{Max-V}/\sigma_1 ) w = 40.11</th>
<th>H</th>
<th>( \sim \text{PROB} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \times ) a. sn’ew.li.ta</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-74.43</td>
<td>0.99</td>
</tr>
<tr>
<td>b. sn’iw.li.ta</td>
<td>-1</td>
<td>-1</td>
<td></td>
<td>-80.24</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. sn’wli.ta</td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>-80.22</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[i] never surfaces anywhere, even in privileged positions, because the faithfulness con-
straint that would prevent it from becoming ATR, \( \text{Id( ATR)} \) and its positional counterpart are weighted at 0.0; whereas the markedness constraint is weighted slightly above 100.
\[(14) \quad \star ATR \text{ outweighs } Id(ATR)+Id(ATR)/\sigma_1.\]

<table>
<thead>
<tr>
<th>/tiqata/</th>
<th>*</th>
<th>/tiqata/</th>
<th>Id(ATR)/\sigma_1</th>
<th>H</th>
<th>~PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. ti.qa.ta</td>
<td>-1</td>
<td>100</td>
<td>0</td>
<td>-100</td>
<td>3e-44</td>
</tr>
<tr>
<td>b. ti.qa.ta</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Finally, in order to find that epenthesis is used to break up \([Ctk]\) clusters, note that PhTAC must outweigh Dep-V.

\[(15) \quad \star PhTAC \text{ outweighs } Dep-V.\]

<table>
<thead>
<tr>
<th>/taqaktk(^h)/</th>
<th>PhTAC (w = 100)</th>
<th>Dep-V (w = 8.26)</th>
<th>H</th>
<th>~PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. ta.qaktk(^h)</td>
<td>-1</td>
<td>-100</td>
<td>1e-40</td>
<td></td>
</tr>
<tr>
<td>b. ta.qa.katk(^h)</td>
<td>-1</td>
<td>-8.26</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

The weightings found by the learner effectively model the phonotactics of Klamath’. The weighting conditions explored in this section (and repeated in (16)) are necessary for any grammar that has the phonotactics of Klamath’, regardless of underlying forms.

\[(16) \quad \text{a. } \star \text{Mid outweighs } Id(\text{HIGH}) \text{ and/or } Max-V. \]
\[ \text{b. } Id(\text{HIGH})+Id(\text{HIGH})/\sigma_1 \text{ outweighs } \star \text{Mid.} \]
\[ \text{c. } Max-V+Max-V/\sigma_1 \text{ outweighs } \star \text{Mid.} \]
\[ \text{d. } \star ATR \text{ outweighs } Id(ATR)+Id(ATR)/\sigma_1. \]
\[ \text{e. } PhTAC \text{ outweighs } Dep-V. \]

4 Stage 2: Learning Underlying Representations

In the next stage, the learner becomes morphologically aware and tries to learn a probability distribution across underlying forms. There are three types of URs that are relevant for our simulations: the concrete URs, /\oe\wi/ and /\oe\i/; the analytically preferred restrictedly abstract UR /\oe\we/ and the never surfacing, very abstract UR, /\oe\wi/. Simulations were run with two sets of these URs:5

- Only the concrete URs- Without abstract URs available, the learner will fail to converge on a single UR, and fail to model the data.
- All abstract URs- The learner prefers to the more restrictedly abstract UR (/\oe\we/) to the never surfacing abstract UR (/\oe\wi/).

5In the simulations implemented below the options of URs was given to the learner, but one could imagine an algorithm that would find the set of URs, similar to the one implemented by Eisenstat (2009) expanded to allow abstract URs. This expansion will greatly expand the search space, so the learner might implement the concepts of local lexica from Merchant & Tesar (2008); Tesar (2014), to incrementally search through URs that differ from surface exponents.
4.1 Concrete URs

If the set of possible URs is restricted to those concrete forms that surface somewhere, the learner fails to settle on any UR for alternating stems. Instead as shown in (17) it puts half the probability in each concrete UR.

(17)  Constraint Weights Learned by second stage with only concrete URs

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Start From (11)</th>
<th>Final</th>
<th>M or F?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID(Hi)</td>
<td>40.1281</td>
<td>43.8452</td>
<td>F</td>
</tr>
<tr>
<td>ID(Hi)/σ₁</td>
<td>40.1281</td>
<td>44.8451</td>
<td>F</td>
</tr>
<tr>
<td>MAX-V</td>
<td>40.1050</td>
<td>44.4934</td>
<td>F</td>
</tr>
<tr>
<td>MAX-V/σ₁</td>
<td>40.1050</td>
<td>44.4934</td>
<td>F</td>
</tr>
<tr>
<td>*MID</td>
<td>74.4312</td>
<td>85.7059</td>
<td>M</td>
</tr>
<tr>
<td>ID(ATR)</td>
<td>0.0000</td>
<td>0.0000</td>
<td>F</td>
</tr>
<tr>
<td>ID(ATR)/σ₁</td>
<td>0.0000</td>
<td>0.0000</td>
<td>F</td>
</tr>
<tr>
<td>DEP-V</td>
<td>8.2561</td>
<td>6.0257</td>
<td>F</td>
</tr>
<tr>
<td>*₁</td>
<td>100.0000</td>
<td>100.0000</td>
<td>M</td>
</tr>
<tr>
<td>PHTAC</td>
<td>100.0000</td>
<td>100.0000</td>
<td>M</td>
</tr>
</tbody>
</table>

The tableaux in (18) and (19) show the grammar with the constraint weightings and UR probabilities learned in (17). The selected URs are shown in the leftmost column, followed by the probability of that input being chosen. Then the harmony for each candidate is calculated for each input. The probability shown is that of each input-output candidate being chosen globally. This means that the probability shown is the probability of the input multiplied by the probability of that output given that input.

(18) /tɛːwi-ta/ receives too much probability.⁶

<table>
<thead>
<tr>
<th>DEEP- /-ta/</th>
<th>P(UR)</th>
<th>Surface</th>
<th>Max-V w = 44.5</th>
<th>DEP-V w = 6.0</th>
<th>H</th>
<th>~Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>/tɛːwi-ta/</td>
<td>.5</td>
<td>a. [tɛːwita]</td>
<td>0</td>
<td>-44.5</td>
<td>2.5e-20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. [tɛwta]</td>
<td>-1</td>
<td>-44.5</td>
<td>2.5e-20</td>
<td></td>
</tr>
<tr>
<td>/tɛw-ta/</td>
<td>.5</td>
<td>c. [tɛwta]</td>
<td>0</td>
<td>-6.0</td>
<td>.00125</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>d. [tɛwta]</td>
<td>-1</td>
<td>-6.0</td>
<td>.00125</td>
<td></td>
</tr>
</tbody>
</table>

Here, (18-a) is near categorically chosen as the output given /tɛːwita/ as the input, as its harmony score is 44.5 better than its competitor, but the probability of the grammar selecting /tɛːwita/-/tɛːwita/ is only .5, because the UR probability is .5. If /tɛw-ta/ is chosen as the input, (18-c) obtains most of the probability. Therefore, the choice of surface form is completely dependent on the choice of UR, with both [tɛwta] and [tɛːwita] receiving near .5 probability. This is an incorrect result, as the learner has never even seen [tɛwta], and always seen [tɛːwita].

⁶Probabilities in tableaux do not add to 1 because of rounding.
A similar result is seen in (19), the grammar outputs /te:wtlk/-[te:wtlk] fifty percent of the time, even though that form is never seen. The correct surface form [te:wtlk] is only chosen half the time. Note here that the two tableaux are contradictory, in order to select [te:wtlk] as the output in (19), more probability must be given to (/te:wi/), but that UR picks the incorrect output form [te:wtla] in tableau (18).

If the learner fails to converge on a grammar that accurately models the data, like here; the learner should be able to open up its search space to allow abstract URs.

### 4.2 Full set of URs

After opening the search space to allow abstract URs, the MaxLex learner is able to model the surface data. The learner settles on the forms with /e/ with close to 1 probability, with the constraint rankings shown in (20).

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Start From (11)</th>
<th>Final</th>
<th>M or F?</th>
<th>UR</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID(HI)</td>
<td>40.12807</td>
<td>46.66452</td>
<td>F</td>
<td>/te:wi/</td>
<td>.000664</td>
</tr>
<tr>
<td>ID(HI)/σ₁</td>
<td>40.12807</td>
<td>43.06800</td>
<td>F</td>
<td>/te:w/</td>
<td>.0000018</td>
</tr>
<tr>
<td>MAX-V</td>
<td>40.10503</td>
<td>42.67768</td>
<td>F</td>
<td>/te:we/</td>
<td>.997917</td>
</tr>
<tr>
<td>MAX-V/σ₁</td>
<td>40.10503</td>
<td>46.27095</td>
<td>F</td>
<td>/te:wi/</td>
<td>.0004002</td>
</tr>
<tr>
<td>*MID</td>
<td>74.43123</td>
<td>85.70757</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID(ADR)</td>
<td>0.000000</td>
<td>0.00003</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID(ADR)/σ₁</td>
<td>0.000000</td>
<td>0.00000</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEP-V</td>
<td>8.25611</td>
<td>8.35856</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*₁</td>
<td>100.00003</td>
<td>99.99998</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH Tac</td>
<td>100.00003</td>
<td>99.99998</td>
<td>M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As the learner has near categorically learned an abstract UR, it is well able to model the data with similar categoricity. The simulation above obtained over .9 probability for all surface forms.

The critical change in constraint weighting learned during this stage is the difference between ID(HI) and MAX-V. As shown above, the weighting learned by the phonotactic grammar has noninitial /e/s undecided between deleting or raising; near a .5 probability for each option (when PH Tac doesn’t interfere). Now, ID(HI) outweighs MAX-V by a margin
of 3.98. Thus, /e/ deletes over 98% of the time as in (21) when phonotactics allow, but raises to [i] over 99% of the time when they don’t (22). Note, these tableaux only show the results for the two most probable URs; since most of the probability is given to /?e:we/, the probability of any of the other URs being chosen is near negligible.

(21) **Id(Hi) outweighs Max-V**

<table>
<thead>
<tr>
<th>DEEP-/ta/ UR</th>
<th>P(UR)</th>
<th>Surface</th>
<th>*Mid 85.7</th>
<th>Id(Hi) 46.7</th>
<th>Max-V 42.7</th>
<th>H</th>
<th>~PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>/?e:we-ta/</td>
<td>.998</td>
<td>a. [?ewita] -1 -1</td>
<td>-132.4</td>
<td>.018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. [?ewta] -1 -1</td>
<td>-128.4</td>
<td>.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. [?eweta] -2 -1</td>
<td>-171.4</td>
<td>2e-19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/?e:wi-ta/</td>
<td>.001</td>
<td>d. [?ewta] -1 -1</td>
<td>-128.4</td>
<td>2e-22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>e. [?ewita] -1 -1</td>
<td>-85.7</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(22) ***Mid and PhTAC+Max-V outweigh Id(hi)**

<table>
<thead>
<tr>
<th>DEEP-/tk/ UR</th>
<th>P(UR)</th>
<th>Surface</th>
<th>PhTAC 100</th>
<th>*Mid 85.7</th>
<th>Id(hi) 46.7</th>
<th>Max-V 42.7</th>
<th>H</th>
<th>~PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>/?e:we-tk/</td>
<td>.998</td>
<td>a. [?ewitk] -1 -1</td>
<td>-132.4</td>
<td>.998</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. [?ewtk] -1 -1</td>
<td>-228.4</td>
<td>2e-42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. [?ewetk] -2 -1</td>
<td>-171.4</td>
<td>1e-17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/?e:wi-tk/</td>
<td>.001</td>
<td>d. [?ewitk] -1 -1</td>
<td>-85.7</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These results confirm that the restrictedly abstract UR is preferred to the otherwise abstract URs, and that abstract URs are learnable by this model. An alternative would be that the learner settled on /i/ as the UR, but as long as /e/ is an option; the learner prefers it.

5 **Why are restrictedly abstract URs preferred?**

The simulation data shows that the restrictedly abstract UR is preferred to the never-surfacing UR. Several factors in the learner lead to this result. To understand those factors, we must understand the distinct constraint rankings that allow each UR to near-categorically surface with the observed forms.

Since all of the possible abstract segments delete when able, Max-V must be weighted below whatever markedness constraint militates against the abstract segment, as shown in (23). In order to prevent the segment from raising or strengthening to [i], Id(High) or Id(ATR) respectively, must outweigh Max-V. Finally as shown in (21-22), if the Ident constraint is also weighted below the markedness constraint (24), and is outweighed by Max-V+PhTAC (25), [i] will occur when phonotactics prevent deletion not [e] or [i].

(23) **Id(Hi) (or Id(ATR)) must outweigh Max-V.**

(24) ***Mid (or *[i]) must outweigh Id(High) (or Id(ATR))**

7These rankings could be refined if the convergence tolerance was made more stringent.
MAX-V+PhTAC must outweigh ID(HIGH) (or ID(ATR)).

The obvious difference between /e/ and /t/ is that a high weighting of ID(HIGH) is required to model the basic phonotactics of the language (repeated in 26) whereas a high weighting of ID(ATR) is not. However, it is not obvious that the weighting condition (26) should have any effect on the weight of the general faithfulness constraint—the grammar could get the same predictions just by putting a lot of weight into the specific faithfulness, and none into the general one.

(26) \[\text{ID(HIGH)} + \text{ID(HIGH)}/\sigma_1 \text{ outweighs } *\text{Mid}\]

Yet, two forces act to prevent the specific faithfulness constraint from receiving all the weight. One, the optimisation function looks at the gradient of the objective function for each possible set of constraint weights; the gradient of the likelihood function for some constraint is equal to the observed violations of that constraint minus the expected violations. Since for any violation of the specific constraint, there is also a violation of the general constraint, any time there are more observed violations of the specific constraint than expected violations (or vice versa) there must also be at least equally many more observed violations of the general constraint for that same data. This makes it very difficult for the learner to raise the weight of the specific constraint while lowering the weight of the general constraint.\(^8\)

Two, the L2 Gaussian prior, used to prevent constraint weights from getting to high, and to keep markedness weighted high and faithfulness low, also has a preference for spreading the weight among constraints (and preferring a general constraint used multiple times) Since the prior for faithfulness constraints is proportional to the sum of the squares of the constraints’ weights, in order to minimise the prior, while also obtaining a weighting condition like that in (26), the learner tries to share the weight between the constraints. Imagine the sum of the constraints had to reach 80 in order to categorically show [e] surfacing in initial syllables: If all the weight is put into one of the constraints, the prior will be proportional to \(80^2=6400\). However, if the weight is shared between the constraints the prior will be proportional to \(40^2+40^2=3200\). In the natural language data from Klamath, this effect is increased, since in order to correctly capture the phonotactic generalisations about [e], three different weighting conditions of this sort must be learned, the one in (26) and the two in (27).

(27) a. \(\text{ID(HIGH)} + \text{ID(HIGH)}/\text{V} \text{ outweighs } *\text{Mid} \) (to protect long [e])

b. \(\text{ID(HIGH)} + \text{ID(HIGH)}/\text{Noun} \text{ outweighs } *\text{Mid} \) (to protect /e/ in noun roots)

If each of these sums must reach a value of 80, the weights that minimise the prior are 60 for ID(HIGH), the general constraint; and 20 for each of the specific constraints. This property is generalizable, given a typologically predicted positional privilege pattern—disjunctive licensing as in Klamath, conjunctive licensing where a segment appears only long in initial syllables of nouns, or any combination of the two—the prior will prefer to put more weight in the general constraint if the segment surfaces faithfully in more positions.

The important result is that the learner selects the UR that is MORE RESTRICTEDLY ABSTRACT. Between /?e:we/ and /?e:wi/ in Klamath’ the more restrictively abstract option

---

\(^8\)This type of argument is more thoroughly explored in the recent work by (Hughto et al., 2015; Staubs et al., 2016; Pater, 2016) in regards to an Agent-Based Model of learning, which makes several different assumptions than the MaxLex model, which will not be covered here; but similarly finds a bias towards general rather than specific constraints in a MaxEnt based framework.
is /ʔeːwe/ because while both have segments that do not appear in any surface exponent of the morpheme–/e/ and /i/-/e/ is able to surface in more positions in Klamath than /i/ is. Thus, the alternation can fill a gap in the lexical distribution of /e/, whereas it would be the only motivation for /i/ appearing in URs.

The prior allows the learner to pick the grammar and lexicon that prefer more restrictedly abstract forms. Both the grammar where /ʔeːwe/ and the grammar where /ʔeːwɪ/ serve as the UR for [ʔeːwɪɾ]-[ʔeːwitʰ] perform equally well on the likelihood of the output data. Since both grammars can be described in HG (Author, in press), the probability of the output data given grammar and lexicon become near categorical regardless which UR is chosen, and therefore both negative log likelihoods approach zero. Thus, the only thing that differs between these two grammar-lexicon pairs is the value the prior gives; basically the sum of the weights of the faithfulness constraints squared. In both grammars, since [e] surfaces everywhere but in noninitial syllables, the weighting condition in (26) must be respected. Therefore ID(HIGH) must be at some non-zero positive value in both grammars.

Now, the learner must check the value of the prior for both the grammar that uses /e/ and that that uses [i]. ID(HIGH) was learned to be 40.13 to get [e]’s distribution by the phonotactic learner in (11). On the other hand, since ID(ATR) is never respected in the surface data, the constraint can be at zero. In order to get a certain /e/ or /i/ to repair to [i] in contexts where it cannot delete, the respective faithfulness constraint must outweigh MAX-V (28) (as seen in the simulation results above in (21)).

\[
\begin{align*}
\text{(28) } & \quad \text{a. For } /e/ \text{ to show } [i] - [\emptyset] \text{ alternation:} \\
& \quad \text{ID(HIGH) outweighs MAX-V} \\
& \quad \text{b. For } /i/ \text{ to show } [i] - [\emptyset] \text{ alternation:} \\
& \quad \text{ID(ATR) outweighs MAX-V}
\end{align*}
\]

Assume without loss of generality MAX-V is constant at 40.11 in both grammars—it must be relatively high ranked, in order to prevent privileged [e] from deleting, (as well as to prevent any other vowels from deleting in a larger constraint set). Now we can find the minimal values for the prior given the (simplified) universal weighting conditions (29) in order to get a grammar with one of the conditions in (28) (let’s assume the constraint must get to at least 45).

\[
\begin{align*}
\text{(29) } & \quad \text{a. MAX-V= 40.11} \\
& \quad \text{b. ID(HIGH)} \geq 40.13 \\
& \quad \text{c. ID(ATR)} \geq 0
\end{align*}
\]

The grammar necessary for /e/ to serve as the UR then must weight ID(HIGH) at 45, and can keep ID(ATR) at 0. For these three constraints, the prior is proportional to \(0^2 + 40.11^2 + 45^2 = 3625\). On the other hand, if /i/ was to serve as the UR, ID(ATR) must reach 45, while the lowest ID(HIGH) can be is 30. Thus, the prior would be proportional to \(30^2 + 40.11^2 + 45^2 = 4525\). Since ID(HIGH)’s minimum weight is above ID(ATR)’s minimum weight, the global minimum must put all of it’s weight into /ʔeːwe/. This example can be fully generalised to any set of weighting conditions where ID(HIGH) must be higher to satisfy phonotactics than ID(ATR), including the weights learned in (11).

The choice to use a UR that surfaces elsewhere in the language is analytically preferable from an information theoretical perspective. By using /e/ in more positions rather than
adding /i/ to the set of segments exploited in URs, the lexicon can be described with a smaller Minimum Description Length, (Rasin & Katzir, in press). The more efficient option is not chosen by any stipulations made in the grammar, but rather is an emergent property of the learning algorithm, using no machinery that is not already in use to enforce a restrictive grammar.

5.1 Further Implications

Further generalizing these results shows that learners tend towards other analytically pleasing results. Imagine a toy language with a similar alternation to that shown in Klamath, where an [i]-[ə] alternation appears in noninitial syllables of verbs. Unlike Klamath, in this toy language a much larger set of possible abstract URs for the alternation that do surface somewhere in the language (30).

In (30), ×s represent environments where a segment or alternation does not appear, and ✔s represent does appear. Each row represents the distribution of segments for a particular position, from most privileged (Noun and initial syllable) to least privileged (Non-noun and non-initial syllable).\(^9\)

\[(30)\] Surface Distribution of different segments in a toy language

<table>
<thead>
<tr>
<th>Noun</th>
<th>σ₁</th>
<th>[e]</th>
<th>[u]</th>
<th>[i]</th>
<th>[i]</th>
<th>[i]-[ə]</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>×</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>✔</td>
<td>✔</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>✔</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✔</td>
</tr>
</tbody>
</table>

In this language, [e] is disjunctively licensed, surfacing whether in nouns or initial syllables; [u] only surfaces in nouns (represented by the checkmarks in the first three rows); [i] is conjunctively licensed, only in nouns and initial syllables; and [i] is never surfacing. All four forms could categorically serve as the UR for the alternation. By looking at the weighting conditions necessary for each of these segments to show the distribution they do, we can find the grammar that minimises the objective function while explaining the alternation. All of these URs for the [i]-[ə] alternation are abstract, because none of them ever surface in the position of the alternation. Due to the actual surface distributions of the segments, [i] is never surfacing and least restrictedly abstract, [i] is more restrictedly abstract than [i]; [u] is more restrictedly abstract than either [i] or [i]; and [e] is most restrictedly abstract (30). This can be seen because each segment appears on the surface in a superset of the positions where each segment to the right of it appear.

All other things being equal, ID(Hi) should have the highest minimum weight of all the constraints considered here. Since ID(Hi)+ID(Hi)/NOUN and ID(Hi)+ID(Hi)/σ₁ both must outweigh the markedness constraint against [e], the prior attempts to put 2/3 of the weight into the general constraint and 1/3 into each specific constraint. For /u/, since only ID(BACK)+ID(BACK)/NOUN must outweigh the markedness, half the weight goes into the

\(^9\)I make no claims about what is more privileged, Non-noun initial syllables or noun noninitial syllables. This would depend on the relative weightings of the faithfulness constraints in the language.
general constraint. For /ɪ/ the weight can be evenly split between (1/3 each) \( \text{Id} \text{(ATR)} \), \( \text{Id} \text{(ATR)} \text{/Noun} \), and \( \text{Id} \text{(ATR)} \text{/σ}_1 \), since only the sum of the three must outweigh the markedness. Finally there is nothing to prevent \( \text{Id} \text{(NASAL)} \) from being weighted at 0. As a result, the learner will choose /e/ as the UR for the alternation; since the distance between its necessary minimum weight is closest to \( \text{Max-V}'s \) (which has a minimum weight at least equal to \( \text{Id} \text{(HIGH)} \)) since [e] must be protected from deletion as well as raising in nouns and initial syllables.

By picking the UR that is most restrictedly abstract, the learner fills the gap in the lexical distribution that it can best fill. This prediction is analytically preferable, but cannot be easily enforced through any grammatical means in a constraint based grammar with Richness of the Base. However, this shows that the choice of analytically satisfying URs is an emergent property of learning, driven by mechanisms already inherent in the learner.

When all things aren’t held equal, one \text{IDENT}(F) \) constraint could be weighted higher than expected simply from the distribution of the abstract URs. This can happen because each \text{IDENT} \) constraint is relevant for a number of segments that all have the same feature. For example, imagine a language with an inventory like Klamath’s [i e a u] with no surface restrictions, and a [u]-[∅] alternation which appears everywhere instead of the [i]-[∅] alternation. Though /o/ never surfaces, it would still be the learners likely choice of abstract UR for the alternation over something like /u/, simply because \( \text{Id} \text{(HIGH)} \) needed to be relatively high ranking in order to protect /e/, throughout the language. Thus, this results in another analytically pleasing emergent result: there is pressure for the segmental inventory used in the lexicon to be symmetric and respect principles of feature economy.

6 Conclusion

This paper has argued that the learnability argument against abstract URs is not sufficient. The same properties that an analyst might look for when picking an abstract UR for an alternation–feature economy, symmetry, minimizing lexical gaps– are in fact emergent biases in a MaxEnt learning framework. If a more restrictedly abstract UR is available, the learner will pick it.

But what happens when a learner has no preferred abstract UR? If there are no distributional reasons to pick one UR over the others–say only never-surfacing URs are available (for which there will usually be many)–the learner should have no reason to prefer /i/ to /ɪ/ or anything else. This is where I suggest the other last-resort strategies belong. If the learner is having this difficulty, it could learn that multiple underlying forms exist for the stem (Pater et al., 2012), or it could clone constraints in order to lexically index an exception (Pater, 2005). This is not to make any claims about how exceptionality is handled, but to show that the data in Klamath is firmly different than the data that involve true lexical exceptionality.

If the best analysis of a phonological pattern is a single underlying form, it is important to know high of a priority that goal is, and in what case does the learner prefer learning anything at all over learning one single form. If these strategies are considered only after forms like Klamath have been learned, it suggests Klamath’s pattern may be more stable than some of these other types of exceptionality, because noise or unlucky learning data distributions could lead to learners biasing one of the many never-surfacing forms slightly above
some other form. The difference in learners leads to different individuals learning different hidden structures for the same data, which may make some different predictions on very low frequency items, or treatment of loan words, or gradient well-formedness judgements.

References


Gouskova, Maria, & Becker, Michael. 2013. Nonce words show that Russian yer alternations are governed by the grammar. NLLT, 31(3), 735–765.


19
Rasin, Ezer, & Katzir, Roni. in press. On evaluation metrics in Optimality Theory. *LI*.


Stanton, Juliet. in press. Learnability shapes typology: the case of the midpoint pathology. *Lg*.


