How abstract is more abstract?[*]

Abstract underlying representations can be learned through emergent feature economy

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 the question of how distinct can these lexically stored phonological forms be from the forms that actually surface has compelled phonologists since Kenstowicz & Kisseberth (1977)

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 [ˈfotəˌgræf]-[fəˈtɑgrəˌfi] (photograph-photography) among other forms in English;

[*] This paper has benefited greatly from comments on previous drafts by Karen Jesney, Hayeon Jang, Reed Blaylock, Roumyana Pancheva, Rachel Walker, Khalil Iskarous, as well as the editors and two anonymous reviewers. The work within has benefited greatly from discussions with the aforementioned people and Paul de Lacy, Brian Smith, Martin Krämer, Giorgio Magri, Jeffrey Heinz, Eric Baković, Stephanie Shih, Jason Riggle, Ezer Rasin, Josh Falk, Brian Hsu, Mairym Llorens, Adam Ussishkin, Joe Pater, Michael Becker, and audiences at OCP12, SSILA 2015, CLS51 and USC PhonLunch.

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. /fotɑgræf/. 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.

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 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, the search space of possible URs grows substantially. However, this space is searchable if the set of URs is structured: given output-drivenness, Tesar (2014) shows those URs with minimal faithfulness violations can be examined first. In some cases, like the Klamath case explored in this paper, number of faithfulness violations is not enough to distinguish between an unbounded number of potential URs, which each differ from the surface form by an equal number of violations.

The choice of UR here can be based on feature economy, first introduced in (de Groot, 1931, 121): "there is a tendency to employ certain accompanying phoneme properties more than once"[1] Under this theory the best UR is the one that makes crucial use of a feature more widely used in the grammar of the language than all other URs. If a feature is contrastive except where the abstract alternation is seen, that same feature can be used to contrast alternating forms from non-alternating ones.

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 2016, a.o.).[2]

This paper claims that the preference to use employ features used elsewhere in the grammar for abstract URs 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). Without any explicit mechanisms

[1] Translation from Clements (2003).

[2] Pater & 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.

built in to drive learning of abstract URs, 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 allomorphy, this paper shows that abstract URs can be learnable, allowing for more alternations to be explained with single URs.

Section 1 will introduce the case study of abstract alternations explored in this paper 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 O'Hara (2015). Klamath has a four vowel place contrast on the surface, [i e a u], and contrastive length for each vowel (Barker, 1964, 1963).

To show the need for abstract URs in Klamath, four types of verb stems will be analyzed with three types of suffixes: /-a/ 'indicative'; /-tkʰ/ 'having been ...ed', and /-tʰa/ (a combination of the indicative and a /-tʰ-/ locative affix meaning 'on').[3]

(3)	*Stem-Type*	*With* /-a/	*With* /-tkʰ/	*With* /-tʰa/	
	/n/-final	[wena]	[wentkʰ]	[wentʰa]	'wear out'
	/C/-final	[slˀeqˀa]	[slˀeqˀatkʰ]	[slˀeqtʰa]	'rust'
	/i/-final	[stupw̥i]	[stupw̥itkʰ]	([stupw̥itʰa])[4]	'has first menstruation'
	Abstract	[ncʰeːwˀa]	[wcʰeːwitkʰ]	[ncʰeːwtʰa]	'break'

Unlike the other types of stems presented in (3), which can be concretely represented as /wen/, /slˀeqˀ/ and /stupw̥i/ respectively, the [i]-[∅] alternation cannot be modeled with concrete URs.

(4) *Verb stems exhibiting* [∅] ~ [*i*] *alternation*

With /-a/	*With* /-tkʰ/
[ʔeːw̥-a] 'is deep'	[ʔeːw̥**i**-tkʰ] 'deep'
[q'eːl'-a] 'acts silly'	[q'eːl'**i**-tkʰ] 'one acting silly'

[3] Conclusive evidence for these underlying forms for these morphemes comes is most clear from their interaction with other suffixes, and these stems and cannot be properly analyzed in the space allotted here, thus for the purposes of this paper I will take these URs as given. However, the simulations search a space of URs for these suffixes (/a/ and /∅/, as well as several vowel initial variants of /-tkʰ/, these URs are settled upon by the successful simulations.

[4] Due to the relative rarity of /i/-final stems, no stems that could pair semantically with the /-t-/ locative affix occur in Barker (1963), for explanatory purposes I show the hypothetical form based on many other suffixes with similar phonotactics, i.e. [sesatwi]-[sesatwitki] 'sell'.

Two potential concrete URs are available for a morpheme like [ʔeːw̥a]-[ʔeːw̥itkʰ]: /ʔeːw̥/ where the [i] is epenthesised, and /ʔeːw̥i/ where the /i/ is deleted. However, neither of these analyses are possible given the rest of the language's phonology. The [i]-[∅] alternation is restricted particularly to noninitial syllables of verbs, whereas the [a] epenthesis process appears across the language (Barker, 1964). There seems to be no phonotactic reason for some stems to show [i] epenthesis and others to show [a] epenthesis, as the phonological environments for both are very similar. Therefore, if /ʔeːw̥/ were the UR, /ʔeːw̥-tkʰ/ would surface as [ʔeːw̥atkʰ] rather than [ʔeːw̥itkʰ].

(5) *Verb stems showing [a] as the default epenthetic vowel*

With /-a/	*With* /-tkʰ/	*With* /-ta/-type	
[sk’aːw̥-a]	[sk’aːw̥-**a**-tkʰ]	[sk’aːw̥-tki]	‘is cold’
[kuw̥-a]	[kuw̥-**a**-tkʰ]	[kuw̥-y’asqs]	‘swells up’

Additional evidence against an [i] epenthesis account comes from a glottalisation alternation present in many of these stems. In Klamath, prevocalic glottal stops coalesce with preceding stops to create glottalised consonants in many contexts, whereas coda glottal stops delete. In many verbs showing the [i]∼[∅] alternation, glottalisation appears on the consonant only when [i] does not. If the [i] was epenthesised here, it could epenthesise after the glottal stop and maximize faithfulness to the underlying form, giving us [ntewˀitkʰ] rather than [ntewitkʰ]. Forms with [a] epenthesis (like [tsaːkˀatkʰ] in (5)) show that when [a] epenthesises, glottalisation is not lost. Only if the vowel exists underlyingly (/ntew**e**ʔ/) can this pattern be modeled.

(6) *Glottalization alternation shows that [i] is not epenthesized*

With /-a/	*With* /-tkʰ/	
[n-te**wˀ**-a]	[n-te**wi**-tkʰ]	‘breaks with round instr.’
[nˀep-sisiː**lˀ**-a]	[nˀep-sisiː**li**-tkʰ]	‘puts a ring on’

Evidence against the /ʔeːi̥/ UR, comes from the many verb stems show nonalternating [i] within this same paradigm, as in (7). When /i/ final stems appear with the /-a/ suffix, deletion of the /a/ occurs to resolve hiatus. Therefore, if /ʔeːw̥i/ were the UR, /ʔeːw̥i-a/ would incorrectly surface as [ʔeːw̥i]. Therefore neither of these concrete URs can model the alternation.

(7) *Verb stems with non-alternating stem-final /i/*

With /-a/	*With* /-tkʰ/	
/stupw̥i-a/→[stupw̥**i**]	[stupw̥**i**-tkʰ]	‘has first menstruation’
/slaːmˀi-a/→[slaːmˀ**i**]	[slaːmˀ**i**-tkʰ]	‘is a widower’

The [i]-[∅] alternation cannot be represented concretely, so a learner or analyst might attempt to represent it abstractly. Importantly for Klamath, [e] appears in all positions (nouns; initial syllables, etc) except where this alternation is found. Since /e/ never appears in this context, but appears elsewhere, it is a particularly good potential UR, whereas a segment like /ɪ/, which never surfaces in Klamath, seems less good. This is because /e/ is more RESTRICTEDLY ABSTRACT than /ɪ/ as defined in (8). [e] has a wider distribution than [ɪ], surfacing in a superset of its positions. Since [ɪ] never surfaces, its distribution is the empty set, and thus any segment that ever surfaces appears in a superset of its positions. Crucially here, a contrast between /e/ and /i/ (or of the feature [±high]) is used in other positions in Klamath, where a contrast between /ɪ/ and /i/ (the feature [±ATR]) is nowhere else used.

(8) 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 /ʔeːw̥e/ and /ʔeːw̥ɪ/ are both abstract morphemes being considered for a [ʔeːw̥]-[ʔeːw̥i]; we look at the distribution of [e] and [ɪ].

b. If the contexts where [e] appear are a superset of those where [ɪ] appears, [ʔeːw̥e] is more restrictedly abstract than [ʔeːw̥ɪ].

Importantly, /ɪ/, and [±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 /ɪ/, it would also be preferred to /ĩ/ or /y/, or even some sort of diacritic approach. A indexed constraint account, (Benua 1997; Alderete 1999; Itô & Mester 1999; Pater 2000, 2005; Coetzee 2009; Becker 2009; Coetzee & Pater 2011; Gouskova & Becker 2013, a.m.o.) for example, would requiring marking relevant morphemes with some exceptional feature used nowhere else in the grammar.

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. An objective function is a function that organizes the search space of possible grammars by quantifying the relative success of a grammar at modeling the data. This objective function is built of two parts. First, the negative log likelihood of the learning data surfacing as observed. This factor quantifies how well the grammar models the learning data; a grammar that has a 100% probability of producing the learning data would receive a zero on this factor, and the lower the probability of producing the data, the higher this factor becomes. Second, a L2 Gaussian prior serves to prevent the constraint weights from getting too large, and to bias markedness constraints high, and faithfulness constraints near 0 (as shown to be necessary for restrictive learning (shown necessary in OT by Smolensky 1996; and by Jesney & Tessier 2011 in HG)). Here the L2 prior is simply a factor of the distance of any constraint from its starting point, resulting in that given two sets of weights that result in the same probability of creating the output data, the learner prefers the one where constraint weights are closer to their initial weights. This factor serves to replicate the ease of learning of the grammar, a grammar that requires more change in constraint weights would require more data for an online learner (like an actual child).

The objective function in (9), used during the first stage of the learner, is no different than the typical objective function used in the MaxEnt literature. That is because at this point the learner is not morphologically aware, so no UR-probabilities need to be included.

(9) *Objective function for first stage of MaxLex*[5]

$$\mathbf{O}_{Ptac}(\mathbf{w}) = \underbrace{-ln(\prod_{o_i \in SurfForms}(\frac{e^{H(/o_i/-[o_i])}}{\sum\limits_{z \in Cand(o_i)} e^{H(/o_i/-[z])}}))}_{Negative\ Log\ Likelihood} + \underbrace{\sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2}}_{L2\ Gaussian\ Prior}$$

The objective function once morphological awareness has kicked in and started the second stage, shown in (10), defines the negative log likelihood of observed data differently. The learner now tries to find the value of the constraint weights ($\mathbf{w}$) and the probability distribution of URs (π) 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 (μ_{ij}) and multiplies this by the probability that u_{ij} is the underlying form for μ_{ij} ($P(u_{ij}|\mu_{ij})$). For this Klamath data, each word is only composed of two morphemes, so only μ_{i1} and μ_{i2} are important for this paper and the simulations within. The objective function for the morphologically aware stage is thus primarily the same as (9), except the likelihood of each input-output mapping is multiplied by the probability of each UR used in the input.

(10) *Objective function for second stage of MaxLex*

$$\mathbf{O}_{Lex}(\mathbf{w}, \pi) =$$

$$\underbrace{-ln(\prod_{\{\mu_{i1},...\mu_{in}\}, o_i \in Data}\left[\sum_{u_{i1} \in UR(\mu_{i1})} P(u_{i1}|\mu_{i1}) ... \sum_{u_{in} \in UR(\mu_{in})} P(u_{in}|\mu_{in})(\frac{e^{H(/u_{i1}...u_{in}/-[o_i])}}{\sum\limits_{z \in Cand(u_{i1}...u_{in})} e^{H(/u_{i1}...u_{in}/-[z])}})\right])}_{Negative\ Log\ Likelihood}$$

[5] Here H represents the harmony score of an input-output candidate, given weights $\mathbf{w}$. $Cand(o_i)$ is the set of candidates given input o_i. c_i returns 100 if the ith constraint is a markedness constraint, and 0 if it is faithfulness. σ_i is a plasticity constant that is set separately for markedness and faithfulness constraints.

$$+ \underbrace{\sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2}}_{L2\ Gaussian\ Prior}$$

I ran a series of simulations (all available in the supplemental materials to this paper) on different training data in order to test the leaners capability to learn the Klamath pattern. Recall from (3) that there are four types of stems which alternate differently with three types of suffixes. I collected a set of 50 verb stems that appear with both /-a/ and /-tk^h/. These are the most common verb suffixes in Klamath; so a large overlap of stems was possible. Exposed to data with just these two suffixes, the learner gave all of the probability for abstract alternating stems like [ʔeːw̥a]-[ʔeːw̥itk^h] to the restrictedly abstract UR /ʔeːw̥e/ above concrete and less restrictedly abstract URs; learning the surface data with .89 probability overall, and over .999 probability for any given form.

In order to handle hiatus resolution with the indicative /-a/ suffix, a set of stem-faithfulness constraints (Beckman, 1998) were necessary. For the purpose of understanding the fundamental emergent principle that leads to this preference for more restrictedly abstract URs, these additional constraints complicate the picture, because they share the burden of the general faithfulness constraints and the bias arises as a cumulative interaction of multiple constraints. In order to simplify the data for the purposes of exposition and highlighting the specific inherent mechanisms that lead to systems preferring to use features they use elsewhere in the grammar, I look instead at stems with the /-tk^h/ suffix and suffixes shaped like /-t^{h}a/ allowing a simulation to be run without constraints addressing hiatus resolution.

Unfortunately, no single suffix of the third shape (like /-t^{h}a/) has a large enough distribution for a large set of verb stems to be selected with all three suffixes or even just with /-tk^h/; though most verb stems pattern with one suffix of this type. For the purposes of the simulation, all of these suffixes would act uniform for all relevant constraints, so I collapsed all of these suffixes into one /-Ca/ suffix.[6] In the following sections, I present the results of this simulation in more detail.

3 Stage 1: Phonotactic Learning of Klamath

In stage 1, the learner is presented with the Klamath data as shown in (11). One set of forms shows the troublesome alternation that is the focus of this inquiry (11-(a)). The other two are necessary to show the learner that the concrete URs cannot serve as the UR for the alternation; (11-(b)) shows that /i/-final stems fail, and (11-(c)) show that consonant final stems fail.

(11)	(a)	([i]-[∅] alternation)	[ʔeːw̥Ca]	[ʔeːw̥itk^h]
	(b)	([i] Nonalternating)	[sn'ewl̥iCa]	[sn'ewl̥itk^h]
	(c)	([a]-Epenthesis)	[sk'aːw̥Ca]	[sk'aːw̥atk^h]

[6] To test a pattern most like actual Klamath, I tested the learner on a dataset with all three types of suffixes, which also correctly learned the more restrictedly abstract forms /ʔew̥e/ for the correct morphemes. This simulation is available in the supplemental materials.

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 (12). 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.

(12) *Constraint Weights Learned by Phonotactic Grammar*

Constraint	Start	Final	M or F?
Id(hi)	50	45.6912869	F
Id(hi)/σ_1	50	45.6912869	F
Max-V	50	45.3103455	F
Max-V/σ_1	50	45.3103455	F
*Mid	50	85.4397388	M
Id(ATR)	50	0.000000	F
Id(ATR)/σ_1	50	0.000000	F
Dep-V	50	8.16611560	F
*i	50	100.0	M
PhTac	50	100.0	M

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/σ_1. The constraint PhTac is a cover constraint here, used to represent overall Klamath phonotactics; it assigns a violation mark to word final clusters of [Ctk^h] (unless the C is [n].).

3.1 Necessary Weightings in HG for Klamath

There are two types of weighting conditions involved in modeling the [i]-[$\emptyset$] 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
- [ɪ] surfaces nowhere
- [Ctk] clusters do not appear word-finally

First, for [e] to not surface in noninitial syllables, at least one of Id(hi) or Max-V must be weighted below *Mid in order to drive some repair of noninitial [e] vowels. In fact, using the weights learned by the phonotactic stage (as in (12)), it is true for both constraints (13). Since [nu.puːsi.Ca] and [nu.puːs.Ca] are so close in weightings, in MaxEnt they each share around half the output probability. Crucially the candidate [nu.pu.se.Ca], with the [e] appearing unlicensed in a nonprivileged position, gets near zero, so it almost never surfaces and is marked with a ✗.

(13) *Mid *outweighs* Id(high) *and/or* Max-V

/nupuːseCa/	*Mid $w = 85.44$	Id(high) $w = 45.69$	Max-V $w = 45.69$	H	~Prob
✗ a. nu.pu.se.Ca	-1			-85.44	3e-18
b. nu.puː.si.Ca		-1		-45.69	.5
c. nu.puːs.Ca			-1	-45.69	.5

[e] surfaces faithfully in initial syllables. Since [e] does not raise in initial syllables, the sum of the weights of Id(high) and Id(high)/σ_1 must be more than the weight of *Mid. To prevent it from deleting, Max-V and Max-V/σ_1 must collectively outweigh *Mid. Both of these results are shown to be learned in (14).

(14) Id(high)+Id(high)/σ_1 *and* Max-V+Max-V/σ_1 *outweigh* *Mid.

/sn'ewl̥iCa/	*Mid $w = 85.44$	Id(high) $w = 45.69$	Id(high)/σ_1 $w = 45.69$	Max-V $w = 45.31$	Max-V/σ_1 $w = 45.31$	H	~Prob
☞ a. sn'ew.l̥i.Ca	-1					-85.44	0.99
b. sn'iw.l̥i.Ca		-1	-1			-91.38	0.003
c. sn'wl̥i.Ca				-1	-1	-90.62	0.003

[ɪ] never surfaces anywhere, even in privileged positions, because the faithfulness constraint that would prevent it from becoming ATR, Id(ATR) and its positional counterpart are weighted at 0.0; whereas the markedness constraint is weighted slightly above 100.

(15) *ɪ *outweighs* Id(ATR)+Id(ATR)/σ_1.

/tɪqaCa/	*ɪ $w = 100$	Id(ATR)/σ_1 $w = 0.0$	Id(ATR) $w = 0.0$	H	∼Prob
a. tɪ.qa.Ca	-1			-100	3e-44
☞ b. ti.qa.Ca		-1	-1	0	1

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

(16) PhTac *outweighs* Dep-V.

/taqaktk^h/	PhTac $w = 100$	Dep-V $w = 8.17$	H	∼Prob
a. ta.qaktk^h	-1		-100	1e-40
☞ b. ta.qa.katk^h		-1	-8.17	1

The weightings found by the learner effectively model the phonotactics of Klamath. The weighting conditions explored in this section (and repeated in (17)) are necessary for any grammar that has the phonotactics of Klamath, regardless of underlying forms.

(17) a. *Mid outweighs Id(high) and/or Max-V.
b. Id(high)+Id(high)/σ_1 outweighs *Mid.
c. Max-V+Max-V/σ_1 outweighs *Mid.
d. *ɪ outweighs Id(ATR)+Id(ATR)/σ_1.
e. PhTac 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, /ʔeːw̥/ and /ʔeːw̥i/; the analytically preferred restrictedly abstract UR /ʔeːw̥e/ and the never surfacing, very abstract UR, /ʔeːw̥ɪ/. Simulations were run with two sets of these URs:[7]

- 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 (/ʔeːw̥e/) to the never surfacing abstract UR (/ʔeːw̥ɪ/).

[7] In 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 (18) it puts half the probability in each concrete UR.

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

Constraint	Start *From (12)*	Final	M or F?
Id(hi)	45.6913	45.756982965	F
Id(hi)/σ_1	45.6913	45.756982965	F
Max-V	45.3103	45.6261506029	F
Max-V/σ_1	45.3103	45.6261505994	F
*Mid	85.4397	85.603224167	M
Id(ATR)	0.0000	0.00000	F
Id(ATR)/σ_1	0.0000	0.00000	F
Dep-V	8.1661	8.16611564629	F
*ɪ	100.0	100.0	M
PhTac	100.0	100.0	M

UR	Prob
/ʔeːw̥i/	.5
/ʔeːw̥/	.5

The tableaux in (19) and (20) show the grammar with the constraint weightings and UR probabilities learned in (18). 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.

(19) *[ʔeːw̥iCa] receives too much probability.*[8]

deep-/ta/ UR	P(UR)	Surface	Max-V $w = 45.63$	Dep-V $w = 8.17$	H	~Prob
/ʔeːw̥i-Ca/	.5	a. [ʔeːw̥iCa]			0	.5
		b. [ʔeːw̥Ca]	-1		-45.63	1.5e-20
/ʔeːw̥-Ca/	.5	c. [ʔeːw̥Ca]			0	.5
		d. [ʔeːw̥aCa]		-1	-8.17	.00028

Here, (19-a) [ʔeːw̥iCa] is near categorically chosen as the output given /ʔeːw̥i-Ca/ as the input, as its harmony score is 44.5 better than its competitor [ʔeːw̥Ca], but the probability of the grammar selecting /ʔeːw̥iCa/-[ʔeːw̥iCa] is only .5, because the UR probability is .5. If /ʔeːw̥-Ca/ is chosen as the input, (19-c) [ʔeːw̥Ca] obtains most of the probability. Therefore, the choice of surface form is completely dependent on the choice of UR, with both [ʔew̥Ca] and [ʔeːw̥iCa] receiving near .5 probability. This is an incorrect result, as the learner has never even seen [ʔeːw̥iCa], and always seen [ʔeːw̥Ca].

[8] Probabilities in tableaux do not add to 1 because of rounding.

(20) *[ʔeːw̥atk^h] receives too much probability.*

deep-/tk^h/ UR	P(UR)	Surface	PhTac $w = 100$	Max-V $w = 45.63$	Dep-V $w = 8.17$	H	∼Prob
/ʔeːw̥i-tk^h/	.5	a. [ʔeːw̥itk^h]				0	.5
		b. [ʔeːw̥tk^h]	-1	-1		-145.63	1e-63
/ʔeːw̥-tk^h/	.5	c. [ʔeːw̥tk^h]	-1			-100	1e-41
		d. [ʔeːw̥atk^h]		-1		-8.17	.5

A similar result is seen in (20), the grammar outputs (20-d) /ʔeːw̥tk^h/-[ʔeːw̥atk^h] fifty percent of the time, even though that form is never seen. The correct surface form (20-a) [ʔeːw̥itk^h] is only chosen half the time. Note here that the two tableaux are contradictory, in order to select (20-a) [ʔeːw̥itk^h], more probability must be given to (/ʔeːw̥i/), but that UR picks the incorrect output form (19-a) [ʔeːw̥iCa] in tableau (19).

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 (21).

(21) *Constraint Weights Learned by second stage with all URs*

Constraint	Start *From (12)*	Final	M or F?
Id(hi)	45.6913	48.2634635165	F
Id(hi)/σ_1	45.6913	43.1253233729	F
Max-V	45.3103	43.008579883	F
Max-V/σ_1	45.3103	48.146719254	F
*Mid	85.4397	85.3995467444	M
Id(ATR)	0.0000	0.00000	F
Id(ATR)/σ_1	0.0000	0.00000	F
Dep-V	8.1661	10.9731942156	F
*ɪ	100.0	100.0	M
PhTac	100.0	100.0	M

UR	Prob
/ʔeːw̥i/	.00000002
/ʔeːw̥/	.0000005
/ʔeːw̥e/	.999999
/ʔeːw̥ɪ/	.0000000008

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 PhTac 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 (22) when phonotactics allow, but raises to [i] over 99% of the time when they don't (23).[9] Note, these tableaux only show the results for the two most probable URs; since most of the probability is given to /ʔeːw̥e/, the probability of any of the other URs being chosen is near negligible.

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

DEEP-/ta/ UR	P(UR)	Surface	*Mid w = 85.4	Id(Hi) w = 48.2	Max-V w = 43.0	H	∼Prob
/ʔeːw̥e-Ca/	.999	a. [ʔeːw̥iCa]	-1	-1		-133.6	.0055
		b. [ʔeːw̥Ca]	-1		-1	-128.4	.9945
		c. [ʔeːw̥eCa]	-2			-170.8	4e-19
/ʔeːw̥i-Ca/	2.4e-8	d. [ʔeːw̥Ca]	-1		-1	-128.4	5e-27
		e. [ʔeːw̥iCa]	-1			-85.4	2.4e-8

(23) *Mid *and* PhTac+Max-V *outweigh* Id(hi)

DEEP-/tk^h/ UR	P(UR)	Surface	PhTac w = 100	*Mid w = 85.4	Id(hi) w = 48.3	Max-V w = 43.0	H	∼Prob
/ʔeːw̥e-tk^h/	.9999	a. [ʔeːw̥itk^h]		-1	-1		-133.7	1
		b. [ʔeːw̥tk^h]	-1	-1		-1	-228.4	6e-158
		c. [ʔeːw̥etk^h]		-2			-170.8	7e-17
/ʔeːw̥i-tk^h/	2.4e-8	d. [ʔeːw̥itk^h]		-1			-85.4	2.4e-8

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 /ɪ/ as the UR, but as long as /e/ is an option; the learner prefers it.

5 Discussion

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. In this section, first the properties underlying the results will be explored; followed by a discussion of their further implications.

5.1 Why are restrictedly abstract URs preferred?

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 (24). 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 (22-23), if the Ident constraint is

[9] These rankings could be refined if the convergence tolerance was made more stringent.

also weighted below the markedness constraint (25), and is outweighed by Max-V+PhTac (26), [i] will occur when phonotactics prevent deletion, rather than the faithful [e] or [ɪ].

(24) Id(hi) (or Id(ATR)) must outweigh Max-V.

(25) *Mid (or *[ɪ]) must outweigh Id(high) (or Id(ATR)

(26) Max-V+PhTac must outweigh Id(high) (or Id(ATR).

The obvious difference between /e/ and /ɪ/ is that a high weighting of Id(high) is required to model the basic phonotactics of the language (repeated in 27) whereas a high weighting of Id(ATR) is not. However, it is not obvious that the weighting condition (27) 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.

(27) Id(high)+Id(high)/σ_1 outweighs *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.[10]

Secondly, the L2 Gaussian prior has a preference for spreading the weight among constraints (and preferring a general constraint used multiple times). Recall this prior is used to prefer grammars where the constraint weights have moved least from their initial weights (markedness high, faithfulness at zero). Since the prior for faithfulness constraints is proportional to the sum of the squares of the constraints' weights ($\sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2}$ when $c_i = 0$), in order to minimise the prior, while also obtaining a weighting condition like that in (27), 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 (27) and the two in (28).

(28) a. (Id(high)+Id(high)/Vː outweighs *Mid (to protect long [eː])
b. (Id(high)+Id(high)/Noun outweighs *Mid (to protect /e/ in noun roots)

[10] 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.

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 /ʔeːw̥e/ and /ʔeːw̥ɪ/ in Klamath the more restrictedly abstract option is /ʔeːw̥e/ because while both have segments that do not appear in any surface exponent of the morpheme– /e/ and /ɪ/–/e/ is able to surface in more positions in Klamath than /ɪ/ is. Thus, the alternation can fill a gap in the lexical distribution of /e/, whereas it would be the only motivation for /ɪ/ appearing in URs.

The prior allows the learner to pick the grammar and lexicon that prefer more restrictedly abstract forms. Both the grammar where /ʔeːw̥e/ and the grammar where /ʔeːw̥ɪ/ serve as the UR for [ʔeːw̥Ca]-[ʔeːw̥itk^h] 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 (27) 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 [ɪ]. Id(high) was learned to be 40.13 to get [e]'s distribution by the phonotactic learner in (12). 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 /ɪ/ to repair to [i] in contexts where it cannot delete, the respective faithfulness constraint must outweigh Max-V (29) (as seen in the simulation results above in (22)).

(29) a. *For /e/ to show [i] − [∅] alternation:*
Id(high) outweighs Max-V
b. *For /ɪ/ to show [i] − [∅] alternation:*
Id(ATR) outweighs Max-V

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 (30) in order to get a grammar with one of the conditions in (29) (let's assume the constraint must get to at least 45).

(30) a. Max-V$=40.11$ b. Id(high)≥ 40.13 c. Id(ATR)≥ 0

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 /ɪ/ 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 /ʔeːw̥e/. 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 (12).

5.2 Further Implications

Further generalizing these results shows that learners tend towards other analytically pleasing results. When given the choice of a variety of abstract underlying representations, a learner will choose to make use of the feature used contrastively in more positions (i.e. with a higher weighted faithfulness constraint associated with it).

Importantly this is not a categorical contrast between restrictedly abstract URs and never surfacing URs, but allows for ordering of URs in terms of how restrictedly abstract they are. If we have a language where [e] appears in all syllables of nouns and initial syllables of words of all categories, and [ɪ] only appears in initial syllables of nouns, /e/ would be a more restrictedly abstract UR for an [i]-[∅] alternation in noninitial syllables of verbs than /ɪ/, causing it to be learned as the UR for this alternation.

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.

However, most faithfulness constraints effect more than one segment in a language. Since the choice of abstract UR is based on the relative weight of each relevant faithfulness constraint, this effect can be seen not just with segments, but with features. 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 in all positions. Though /o/ never surfaces, it would still be the learner's likely choice of abstract UR for the alternation over something like /ʊ/, simply because Id(high) needs to be relatively high ranking in order to protect /e/ from raising to [i]. Thus, this results in another analytically pleasing emergent result: learners prefer to minimize the number of contrastive features in their language when learning their lexicon, resulting in a more symmetric inventory of underlying segments than predicted by chance.

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. Thus the set of possible URs for a morpheme can include the surface exponents themselves, amalgams of the surface exponents (concrete URs), and abstract URs.

But what happens when a learner has no preferred abstract UR? Many Slavic languages show exceptional 'yer' vowels that delete when phonotactically able (Jarosz, 2005; Gouskova

& Becker, 2013). However, unlike Klamath, there are no vowels in complementary distribution with this alternation. If there are no distributional reasons to pick one UR over the others—in Slavic languages, only never-surfacing URs are available (for which there will usually be many)–the learner should have no reason to prefer /ɪ/ to /ĩ/ 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 how 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

Albright, Adam. 2002 (June). *The Identification of Bases in Morphological Paradigms.* Ph.D. thesis, University of California, Los Angeles, Los Angeles.

Alderete, John. 1999. *Morphologically governed accent in Optimality Theory.* Ph.D. thesis, University of Massachusetts Amherst, Amherst.

Alderete, John. 2008. Using learnability as a filter on factorial typology: A new approach to Anderson and Browne's generalization. *Lingua*, **118**, 1177–1220.

Alderete, John, Brasoveanu, Adrian, Merchant, Nazarré, Prince, Alan, & Tesar, Bruce. 2005. Contrast Analysis Aids the Learning of Phonological Underlying Forms. *Pages 34–42 of:* Alderete, John, Han, Chung-hye, & Kochetov, Alexei (eds), *WCCFL 24.* Somerville, MA: Cascadilla Proceedings Project.

Allen, Blake, & Becker, Michael. 2015. *Learning alternations from surface forms with sublexical phonology.* Ms. available (February 2016) at http://ling.auf.net/lingbuzz/002503.

Baković, Eric. 2009. Abstractness and motivation in phonological theory. *Studies in Hispanic and Lusophone Linguistics*, **2**.

Barker, M.A.R. 1963. *Klamath Dictionary.* Berkeley and Los Angeles: University of California Press.

Barker, M.A.R. 1964. *Klamath Grammar.* Berkeley and Los Angeles: University of California Press.

BECKER, MICHAEL. 2009. *Phonological Trends in the Lexicon: The Role of Constraints.* Ph.D. thesis, University of Massachusetts Amherst, Amherst, MA.

BECKMAN, JILL N. 1998. *Positional Faithfulness.* Ph.D. thesis, University of Massachusetts Amherst, Amherst.

BENUA, LAURA. 1997. *Transderivational identity: phonological relations between words.* Ph.D. thesis, University of Massachusetts Amherst, Amherst.

BOWERS, DUSTIN. 2015. *A System for Morphophonological Learning and its Consequences for Language Chagne.* Ph.D. thesis, University of California Los Angeles.

CLEMENTS, G. N. 2003. Feature economy in sound systems. *Phonology*, **20**, 287–333.

COETZEE, ANDRIES. 2009. Learning indexical indexation. *Phonology*, **26**, 109–145.

COETZEE, ANDRIES, & PATER, JOE. 2011. The place of variation in phonological theory. *Pages 401–431 of:* GOLDSMITH, JOHN A., RIGGLE, JASON, & YU, ALAN (eds), *The Handbook of Phonological Theory*, 2nd ed edn. Malden, MA: Blackwell.

DE GROOT, A. W. 1931. Phonologie und Phonetik als Funktionswissenschaften. *Travaux du Cercle Linguistique de Prague*, **4**, 116–147.

EISENSTAT, SARAH. 2009. *Learning underlying forms with MaxEnt.* M.Phil. thesis, Brown University.

FLORA, JO ANN. 1974. *Palauan phonology and morphology.* Ph.D. thesis, University of Califonia San Diego, San Diego, California.

GOLDWATER, SHARON, & JOHNSON, MARK. 2003. Learning OT constraint rankings using a Maximum Entropy model. *In: Proceedings of the Workshop on Variation within Optimality Theory.* Stockholm University.

GOUSKOVA, MARIA, & BECKER, MICHAEL. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *NLLT*, **31**(3), 735–765.

HAYES, BRUCE. 2004. Phonological acquisition in Optimality Theory: the early stages. *In:* KAGER, RENÉ, PATER, JOE, & ZONNEVELD, WIM (eds), *Fixing Priorities: Constraints in Phonological Acquisition.* Cambridge: Cambridge University Press.

HAYES, BRUCE, & WILSON, COLIN. 2008. A maximum entropy model of phonotactics and phonotactic learning. *LI*, **39**, 379–440.

HEINZ, JEFFREY. 2010. Learning Long-Distance Phonotactics. *LI*, **41**, 623–661.

HOCKETT, CHARLES F. 1955. *A manual of phonology.* Baltimore, Maryland: Waverly Press.

Hughto, Coral, Pater, Joe, & Staubs, Robert. 2015 (April). *Grammatical agent-based modeling of typology.* Paper presented at the GLOW Workshop on Computation, Learnability and Phonological Theory, slides at http://blogs.umass.edu/pater/files/2011/10/hughto-pater-staubs-glow.pdf.

Itô, Junko, & Mester, Armin. 1999. The phonological lexicon. *Pages 62–100 of:* Tsujimura, N. (ed), *The Handbook of Japanese Linguistics.* Malden, MA: Blackwell.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. *Pages 467–479 of: Architectures, rules, and preferences: A Festschrift for Joan Bresnan.* CSLI.

Jäger, Gerhard, & Rosenbach, Anette. 2006. The winner takes it all - almost: cumulativity in grammatical variation. *Linguistics*, **44**, 937–971.

Jarosz, Gaja. 2005. Polish Yers and the finer structure of output-output correspondence. *In: BLS 31.*

Jarosz, Gaja. 2006 (October). *Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory.* Ph.D. thesis, John Hopkins University, Baltimore, Maryland.

Jesney, Karen, & Tessier, Anne-Michelle. 2011. Biases in Harmonic Grammar: The road to restrictive learning. *NLLT*, **29**.

Kenstowicz, Michael, & Kisseberth, Charles. 1977. *Topics in Phonological Theory.* New York: Academic Press.

Kenstowicz, Michael, & Kisseberth, Charles. 1979. *Generative Phonology: Description and Theory.* New York: Academic Press.

Kiparsky, Paul. 1968. *How abstract is phonology?* Indiana University Linguistics Club.

Kraft, Dieter. 1988. *A software package for sequential quadratic programming.* Tech. rept. DFVLR-FB 88-28. German Aerospace Center - Institute for Flight Mechanics, Koln, Germany.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. *Phonology*, 213–269.

Martinet, André. 1968. Phonetics and linguistic evolution. *Pages 464–487 of:* Malmberg, Bertil (ed), *Manual of phonetics.*

Merchant, Nazarré. 2008. *Discovering underlying forms: Contrast pairs and ranking.* Ph.D. thesis, Rutgers University, New Brunswick, NJ.

Merchant, Nazarré, & Tesar, Bruce. 2008. Learning Underlying forms by searching restricted lexical subspaces. *In: CLS 41.*

O'Hara, Charlie. 2015. Positionally Abstract Underlying Representations in Klamath. *In: The Proceedings of CLS 51.*

Pater, Joe. 2000. Non-uniformity in English secondary stress: the role of ranked and lexically specific constraints. *Phonology*, **17**, 237–274.

Pater, Joe. 2005. Learning a stratified grammar. *Pages 482–492 of:* Brugos, Alejna, Clark-Cotton, Manuella R., & Ha, Seungwan (eds), *Proceedings of the 29th Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.

Pater, Joe. 2016 (February). *Learning in typological prediction: Grammatical agent-based modeling.* Presented at BLS 42.

Pater, Joe, & Staubs, Robert. 2013. *Feature economy and iterated grammar learning.* Presented at the 21st Manchester Phonology Meeting.

Pater, Joe, Staubs, Robert, Jesney, Karen, & Smith, Brian. 2012. Learning probabilities over underlying representations. *Pages 62–71 of: Proceedings of the twelfth meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology and Morphology.*

Schane, Sanford. 1974. How abstract is abstract? *Pages 297–314 of: CLS Natural Phonology Parasession.*

Smolensky, Paul. 1996. *The Initial State and 'Richness of the Base' in Optimality Theory.* Tech. rept. John Hopkins University.

Stanton, Juliet. 2016. Learnability shapes typology: the case of the midpoint pathology. *Lg.*

Staubs, Robert. 2014. *Computational modeling of learning biases in stress typology.* Ph.D. thesis, University of Massachusetts Amherst, Amherst.

Staubs, Robert, & Pater, Joe. in press. Learning serial constraint-based grammars. *In:* McCarthy, John J., & Pater, Joe (eds), *Harmonic Grammar and Harmonic Serialism.* Equinox.

Staubs, Robert, Culbertson, Jennifer, Hughto, Coral, & Pater, Joe. 2016 (January). *Grammar and learning in syntactic and phonological typology.* Poster Presented at LSA Annual Meeting 2016.

Tesar, Bruce. 2014. *Output-Driven Phonology.* New York: Cambridge University Press.

Tessier, Anne-Michelle. 2007. *Biases and Stages in Phonological Acquisition.* Ph.D. thesis, University of Massachusetts Amherst.

Wilson, Colin. 2006. Learning phonology with a substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science*, **30**(5), 945–982.