Phonology **34** (2017) 325–345. © Cambridge University Press 2017
doi:10.1017/S0952675717000161

How abstract is more abstract? Learning abstract underlying representations[*]

Charlie O'Hara
University of Southern California

This paper presents a Maximum Entropy learner of grammars and lexicons (MaxLex), and demonstrates that MaxLex has an emergent preference for minimally abstract underlying representations. In order to keep the weight of faithfulness constraints low, the learner attempts to fill gaps in the lexical distribution of segments, making 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. MaxLex is shown to be successful in learning certain abstract underlying forms through simulations based on the [i] ~ [∅] alternation in Klamath verbs. The Klamath pattern cannot be represented or learned using concrete underlying representations, 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 these lexically stored phonological forms can be from the forms that actually surface has been of interest to 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

[*] E-mail: CHARLESO@USC.EDU.

This paper has benefited greatly from comments on previous drafts by Karen Jesney, Reed Blaylock, Khalil Iskarous, Hayeon Jang, Roumyana Pancheva, Rachel Walker, as well as the editors and two anonymous reviewers. The work here has benefited greatly from discussions with the abovementioned people and Eric Baković, Michael Becker, Paul de Lacy, Josh Falk, Jeffrey Heinz, Brian Hsu, Martin Krämer, Giorgio Magri, Mairym Lloréns Monteserín, Joe Pater, Ezer Rasin, Jason Riggle, Stephanie Shih, Brian Smith and Adam Ussishkin, as well as audiences at OCP 12, SSILA 2015, CLS 51 and the USC PhonLunch.

The stimuli referred to in the paper are available as online supplementary materials at https://doi.org/10.1017/S0952675717000161.

too strong; some morphemes have alternants that are not predictable in this way. Consider [ˈfotəˌgræf] *photograph* ~ [fəˈtɑgrəˌfi] *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 (Flora 1974, Schane 1974) and elsewhere. However, both of these patterns can be modelled 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, as in (1) (definitions adapted from Kenstowicz & Kisseberth 1979, Baković 2009 and Bowers 2015).

(1) a. A CONCRETE UR is one such that each feature in the UR appears in at least one of its surface exponents.

b. An ABSTRACT UR is one such that 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 (e.g. Kiparsky 1968, Albright 2002, Allen & Becker 2015) that languages should not make use of abstract URs, citing the difficulties they present for the learner. Each morpheme has a limited number of possible concrete URs, correlated to the number and length of surface alternants. 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 that those URs with minimal faithfulness violations can be examined first. In some cases, like the Klamath case explored in this paper, the 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' [translation from Clements 2003]. Under this theory, the best UR is the one that makes crucial use of a feature that is contrastive more often in the grammar of the language than other URs. If a feature is contrastive except where the abstract alternation is seen, that same feature can be used to contrast alternating forms with non-alternating ones.

Feature-economic notions have a long pedigree in phonological analysis (see especially Hockett 1955 and Martinet 1968, and Clements 2003: §1.5 for a historical overview). However, since optimality-theoretic and other related constraint-based grammars tend to have no restrictions on the input and no morpheme-structure constraints, there is no explicit place for feature economy in the grammar. Yet grammar is not the only way to account for language universals and tendencies – recent work has begun to show the power of learnability to restrict typology, through emergent properties (e.g. Alderete 2008, Heinz 2010, Pater & Staubs 2013, Stanton 2016).[1]

This paper claims that the preference for employing features used elsewhere in the grammar for abstract URs emerges naturally out of a learner like that typically used in the MaxEnt learning literature (e.g. Goldwater & Johnson 2003, Jäger & Rosenbach 2006, Wilson 2006, Jäger 2007, Hayes & Wilson 2008). Without any explicit mechanisms built in to drive learning of abstract URs, this emergent bias prefers URs that minimise gaps in the lexical distribution of segments without any direct pressure to do so. While the question of where the line should be drawn between a single UR and allomorphy is largely undecided, this paper shows that abstract URs can be learnable, allowing for a larger class of alternations to be explained with single URs.

§1 introduces the case study of abstract alternations in Klamath explored in this paper, and shows analytically that a concrete UR will fail, and thus that an abstract UR may be useful. The MaxLex learning model is presented in §2, and in §3 I show the results of phonotactic learning and the weighting conditions which are necessary regardless of the selected UR. In §4 the simulation results show that a MaxLex learner learns abstract URs, in particular the most restrictedly abstract UR. §5 offers explanations as to why the learner prefers certain types of abstract URs to others. Finally, in §6, I conclude, and consider further questions.

1 Abstractness in Klamath

In this section, I will sketch out data from a vowel alternation in Klamath (Penutian; Southern Oregon) that motivates the learning of some abstract URs. All Klamath data comes from Barker (1963, 1964) (see O'Hara 2015 for more details of this analysis). Klamath has four surface vowels, [i e a u], with contrastive length for each vowel (Barker 1963, 1964).

To show the need for abstract URs in Klamath, four types of verb stems will be analysed, with three types of suffixes: /-a/ (INDICATIVE), /-tk^{h}/ 'having been VERB-ed' and /-t^{h}a/ (a combination of the indicative and a /-t^{h}-/ locative affix meaning 'on').[2]

[1] 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.

[2] Conclusive evidence for the underlying forms for these morphemes is clearest from their interaction with other suffixes, and these stems cannot be properly analysed in the space available here. For the purposes of this paper I will therefore take the URs as given. However, the simulations search a space of URs for these suffixes (/a/ and /∅/, as well as several vowel-initial variants of /-tk^{h}/), and the URs are settled upon by the successful simulations.

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 a hypothetical form based on many other suffixes with similar phonotactics, e.g. [salɢi] 'miss each other' ~ [salɢit^{h}a] 'suspect each other'.

(2) *stem type*	__/-a/	__/-tk^{h}/	__/-t^{h}a/	
/n/-*final*	[wena]	[wentk^{h}]	[went^{h}a]	'wear out'
C-*final*	[sl^{ʔ}eq^{ʔ}a]	[sl^{ʔ}eq^{ʔ}atk^{h}]	[sl^{ʔ}eqt^{h}a]	'rust'
/i/-*final*	[stupw̥i]	[stupw̥itk^{h}]	[stupw̥it^{h}a]	'has first menstruation'
abstract	[nc^{h}eːw^{ʔ}a]	[wc^{h}eːwitk^{h}]	[nc^{h}eːwt^{h}a]	'break'

Unlike the other types of stems presented in (2), which can be concretely represented as /wen/, /sl^{ʔ}eq^{ʔ}/ and /stupw̥i/ respectively, the [i] ~ [∅] alternation exemplified in (3) cannot be modelled with concrete URs.

(3) *Verb stems showing* [i] ~ [∅] *alternation*

__/-a/		__/-tk^{h}/	
[∅] *surfaces*		[i] *surfaces*	
[ʔeːw̥a]	'is deep'	[ʔeːw̥itk^{h}]	'deep'
[q^{ʔ}eːl^{ʔ}a]	'acts silly'	[q^{ʔ}eːl^{ʔ}itk^{h}]	'one acting silly'

Two potential concrete URs are available for a morpheme like [ʔeːw̥a] ~ [ʔeːw̥itk^{h}] in (3): /ʔeːw̥/, with [i]-epenthesis, and /ʔeːw̥i/, with /i/-deletion. However, neither of these analyses is possible, given the rest of the language's phonology. In particular, the [i] ~ [∅] alternation is restricted to non-initial syllables of verbs, whereas the [a]-epenthesis process in (4) appears across the language (Barker 1964). There seems to be no phonotactic reason for some stems to show [i]-epenthesis and others [a]-epenthesis, as the phonological environments for both are very similar. Therefore, if /ʔeːw̥/ were the UR, /ʔeːw̥-tk^{h}/ would surface as [ʔeːw̥atk^{h}] rather than [ʔeːw̥itk^{h}].

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

__/-a/	__/-tk^{h}/	__/-Ca/	
[∅] *surfaces*	[a] *surfaces*	[∅] *surfaces*	
[taq^{ʔ}a]	[taq^{ʔ}atk^{h}]	[daqn^{ʔ}i]	'sharp-edged'
[kuw̥a]	[kuw̥atk^{h}]	[kuw̥j^{ʔ}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 [i] was epenthesised here, it could appear after the glottal stop and maximise faithfulness to the underlying form, giving [ntew^{ʔ}itk^{h}] rather than [ntewitk^{h}]. Forms with [a]-epenthesis (like [tsaːk^{ʔ}atk^{h}] in (4)) show that when [a] epenthesises, glottalisation is not lost. Only if the vowel exists underlyingly (/nteweʔ/) can this pattern be modelled, as in (5).

(5) *Glottalisation alternation showing that* [i] *is not epenthesised*

__/-a/	__/-tkʰ/	
[ˀ] *surfaces*	[ˀ] *does not surface*	
[ntewˀa]	[ntewitkʰ]	‘breaks with round (INSTR)’
[nˀepsisiːlˀa]	[nˀepsisiːlitkʰ]	‘puts a ring on’

Evidence against the /ʔeːw̥i/ UR comes from the many verb stems that show non-alternating [i] within this same paradigm, as in (6). When /i/-final stems appear with the /-a/ suffix, deletion of the /a/ occurs to resolve hiatus. If /ʔeːw̥i/ were the UR, /ʔeːw̥i-a/ would incorrectly surface as [ʔeːw̥i]. Therefore, neither of these concrete URs can model the alternation.

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

__/-a/	__/-tkʰ/	
/-a/ *deletes*	[i] *surfaces*	
/stupw̥i-a/ → [stupw̥i]	[stupw̥itkʰ]	‘has first menstruation’
/slaːmˀi-a/ → [slaːmˀi]	[slaːmˀitkʰ]	‘is a widower’

The [i] ~ [∅] alternation cannot be represented concretely, so a learner 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 (7). Since [ɪ] never surfaces, its distribution is the empty set, and so any segment that surfaces anywhere appears in a superset of its positions. Thus [e] has a wider distribution than [ɪ]. Crucially here, a contrast between /e/ and /i/ (or of the feature [±high]) is used in other positions in Klamath, while a contrast between /ɪ/ and /i/ (involving the feature [±ATR]) is used nowhere else.

(7) Consider two abstract URs for some alternating morpheme, $x = /x_1 \ldots x_n/$ and $y = /y_1 \ldots y_n/$. Since both are abstract, there must exist some segment in x, say x_i, such 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 y_j in the entire system of the language.

a. If /ʔeːw̥e/ and /ʔeːw̥ɪ/ are both abstract morphemes being considered for a [ʔeːw̥] ~ [ʔeːw̥i] alternation, 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 stand-ins for any means of marking the UR as exceptional that makes use of a feature that is never contrastive on the surface. If /e/ is preferred to /ɪ/ by the learner, it would also be preferred to /ĩ/, /y/ or the diacritic representation /i_1/. An indexed constraint account (Itô & Mester 1999, Pater 2000, Coetzee & Pater 2011, Gouskova & Becker 2013, among others), for example, would require marking relevant morphemes with some exceptional feature used nowhere else in the grammar.

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

2 MaxLex

The MaxEnt learner of grammars and lexicons (MaxLex) discussed in this paper makes use of several theoretical assumptions common in previous learners. MaxLex uses a MaxEnt Harmonic Grammar as its model of constraint interaction. Following previous literature (Hayes 2004, Jarosz 2006, Jesney & Tessier 2011, Tesar 2014), the learner has two stages, first a phonotactic stage, 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. MaxLex differs from many other MaxEnt learners of underlying forms, which use UR constraints (Eisenstat 2009, Pater *et al*. 2012, Staubs & Pater 2016), in that it learns a probability distribution across a set of possible URs, just as Jarosz's (2006) Maximum Likelihood learner of lexicons and grammars does in an OT framework.

The learning algorithm implements a batch learner that minimises an objective function. An objective function is a function that organises the search space of possible grammars by quantifying the relative success of a grammar at modelling the data. This objective function is built of two parts. First, the negative log likelihood of the learning data surfacing as observed is established. This factor quantifies how well the grammar models the learning data; a grammar that has a 100% probability of producing the learning data would be assigned a zero for this factor, and the lower the probability of producing the data, the higher this factor becomes. Second, a L2 Gaussian prior prevents the constraint weights growing indefinitely, as well as modelling an initial bias of higher weights for markedness constraints than faithfulness constraints (in both OT (Smolensky 1996) and HG (Jesney & Tessier 2011)). Here the L2 prior is simply a factor of the distance of any constraint from its starting point, resulting in the learner preferring the grammar where constraint weights are closer to their initial weights, given two grammars that model the learning data equally well. 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 a child).

The objective function in (8), used during the first stage of the learner, does not differ from 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.[3]

(8) *Objective function for the first stage of MaxLex*

$$\mathbf{O}_{Phonotac}(\mathbf{w}) = \underbrace{-ln(\prod_{o_i \in SurfForms} (\frac{e^{\mathcal{H}(/o_i/\rightarrow[o_i])}}{\sum_{z \in Cand(o_i)} e^{\mathcal{H}(/o_i/\rightarrow[z])}}))}_{\text{negative log likelihood}} + \underbrace{\sum_{w_i \in \mathbf{w}} \frac{(w_i - c_i)^2}{\sigma_i^2}}_{\substack{\text{L2 Gaussian}\\ \text{prior}}}$$

The objective function required once morphological awareness has kicked in and started the second stage, which is shown in (9), defines the negative log likelihood of observed data in a different way. The learner now tries to find the value of the constraint weights (**w**) and the probability distribution of URs (π) that maximise the likelihood of the observed data. The critical changes here are that, instead of just looking at the surface forms of the data, o_i, the learner is now aware of a set of n morphemes which together build the word; $\{\mu_{i1} \ldots \mu_{in}\}$. The Harmony and Candidate functions are now evaluated over concatenated strings of possible URs for morphemes, $(u_{i1} \ldots u_{in})$. Since at this stage the learner must find the likelihood of the data given any possible UR, it 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 the simulations in this paper. The objective function for the morphologically aware stage is thus primarily the same as (8), except the likelihood of each input–output mapping is multiplied by the probability of each UR used in the input.

(9) *Objective function for the second stage of MaxLex*

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

$$\underbrace{-ln(\prod_{\substack{\{\mu_{i1} \ldots \mu_{in}\}\\ o_i \in Data}} [\sum_{u_{i1} \in UR\{\mu_{i1})} p(u_{i_1}|\mu_{i_1}) \ldots \sum_{u_{in} \in UR(\mu_{in})} p(u_{in}|\mu_{in}) (\frac{e^{\mathcal{H}(/u_{i1} \ldots u_{in}/\rightarrow[o_i])}}{\sum_{z \in Cand(\mu_{in} \ldots \mu_{in})} e^{\mathcal{H}(/u_{i1} \ldots u_{in}/\rightarrow[z])}})])}_{\text{negative log likelihood}}$$

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

[3] Here $\mathcal{H}$ represents the harmony score of an input–output candidate, given weights **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.

I ran a series of simulations (available in the online supplementary materials) on different training data, in order to test the learner's ability to acquire the Klamath pattern. Recall from (2) 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 /-tkʰ/. 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 categorically preferred the restrictedly abstract UR /ʔeːw̥e/ to the concrete and less restrictedly abstract URs for abstract alternating stems like [ʔeːw̥a] ~ [ʔeːw̥itkʰ], learning the surface data with a probability of 0.89 overall, and greater than 0.999 for any given form.

In order to handle hiatus resolution with the indicative /-a/ suffix, a set of stem-faithfulness constraints (Beckman 1998) was 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, not mechanically or mathematically, but explanatorily. Since the stem-faithfulness constraints share some of the burden of the general faithfulness constraints, their weights would need to be considered in all tableaux and calculations below. Instead, for the rest of the paper, I train the learner on the basis of forms with the /-tkʰ/ suffix and suffixes that act like /-tʰa/ phonotactically, as in (2), allowing the simulation to be run without constraints addressing hiatus resolution.

Unfortunately, no single suffix that behaves like /-tʰa/ has a large enough distribution for a large set of verb stems to be selected with all three suffixes or even just with /-tkʰ/; though most verb stems pattern with one suffix of this type. For the purposes of the simulation, all of these suffixes would behave uniformly for all relevant constraints, so I collapsed all of these suffixes into one /-Ca/ suffix.[4] In the following sections, I present the results of this simulation in more detail.

3 Stage 1: phonotactic learning of Klamath

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

(10)	a.	[i] ~ [∅] *alternation*	[ʔeːw̥Ca]	[ʔeːw̥itkʰ]
	b.	*non-alternating* [i]	[snˀewl̥iCa]	[snˀewl̥itkʰ]
	c.	[a]-*epenthesis*	[skˀaːw̥Ca]	[skˀaːw̥atkʰ]

[4] To test a pattern closely resembling 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 online supplementary materials.

The Candidate function takes the input, and returns 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, so as to prevent infinite epenthesis and allow us to get by with just one cluster constraint. This restricts Gen to consider only candidates that differ with respect to 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 Sequential Least Squares Programming (Kraft 1988), an optimisation algorithm implemented in SciPy, in order to find the constraint weights (and at stage 2 the lexical probabilities) that lead to the minimal objective function.

The constraints used in the simulations are shown in Table I, with the initial weights of each constraint all set uniformly at 50, as well as the weights output by the learner after convergence at this stage. Markedness or faithfulness is also indicated, in order to show the bias of each constraint: 100 for markedness constraints, and 0 for faithfulness constraints.

		initial	final
Ident[high]	faithfulness	50.00	45.69
Ident[high]$_{\sigma1}$	faithfulness	50.00	45.69
Max(V)	faithfulness	50.00	45.31
Max(V)$_{\sigma1}$	faithfulness	50.00	45.31
*Mid	markedness	50.00	85.44
Ident[ATR]	faithfulness	50.00	0.00
Ident[ATR]$_{\sigma1}$	faithfulness	50.00	0.00
Dep(V)	faithfulness	50.00	8.17
*ɪ	markedness	50.00	100.00
Phonotac	markedness	50.00	100.00

Table I

Constraint weights learned by phonotactic grammar.

Most constraints act as expected; a few are worth drawing attention to. For the purposes of this paper, positions in regard to positional faithfulness constraints are defined on the input. In order to model positional asymmetries in deletion patterns we must include Max(V)$_{\sigma1}$. The constraint Phonotac is a cover constraint, used to represent overall Klamath phonotactics; it assigns a violation to word-final clusters of [Ctk^h] (unless C is [n]).

3.1 Necessary weightings in HG for Klamath

There are two types of weighting conditions involved in modelling the [i] ~ [∅] 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, allowing 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 hold for the phonotactics of Klamath are given in (11).

(11) a. [e] does not surface in non-initial syllables.
b. [ɪ] does not surface anywhere.
c. [Ctk] clusters do not appear word-finally.

First, for [e] to not surface in non-initial syllables, at least one of Ident [high] or Max(V) must be weighted below *Mid, in order to drive some repair of non-initial [e] vowels. In fact, using the weights learned at the phonotactic stage (as in Table I), it is true for both constraints, as shown in (12). Since [nu.puːsi.Ca] and [nu.puːs.Ca] are so close in weighting, in MaxEnt they each have around half the output probability. Crucially, the weight of [nu.pu.se.Ca], with [e] appearing unlicensed in a non-privileged position, approaches zero, so this candidate almost never surfaces, and is marked with 👎.

(12)

nupuːseCa	*Mid 85.44	Id[high] 45.69	Max(V) 45.69	$\mathcal{H}$	~p
👎 a. nu.pu.se.Ca	−1			−85.44	3e-18
b. nu.puː.si.Ca		−1		−45.69	0.5
c. nu.puːs.Ca			−1	−45.69	0.5

[e] surfaces faithfully in initial syllables. Since [e] is not raised in initial syllables, the sum of the weights of Ident[high] and Ident[high]$_{\sigma 1}$ must be more than the weight of *Mid. To prevent it from deleting, Max(V) and Max(V)$_{\sigma 1}$ must collectively outweigh *Mid. Both of these weighting conditions are shown to be learned in (13).

(13)

snˀewl̥iCa	*Mid 85.44	Id[hi] 45.69	Id[hi]$_{\sigma 1}$ 45.69	Max(V) 45.31	Max(V)$_{\sigma 1}$ 45.31	$\mathcal{H}$	~p
☞ a. snˀew.l̥i.Ca	−1					−85.44	0.990
b. snˀiw.l̥i.Ca		−1	−1			−91.38	0.003
c. snˀwl̥i.Ca				−1	−1	−90.62	0.003

[ɪ] never surfaces anywhere, even in privileged positions, because the faithfulness constraints that would prevent it from becoming [+ATR], i.e. Ident[ATR] and its positional counterpart, are weighted at 0, whereas the markedness constraint is weighted at 100, as in (14).

(14)

tɪqaCa	*ɪ 100	Id[ATR]$_{\sigma1}$ 0	Id[ATR] 0	$\mathcal{H}$	~p
a. tɪ.qa.Ca	−1			−100	3e-44
☞ b. ti.qa.Ca		−1	−1	0	1

Finally, note that Phonotac must outweigh Dep(V), as in (15), to ensure that epenthesis is used to break up [Ctk] clusters.

(15)

taqaktk^h	Phonotac 100	Dep(V) 8.17	$\mathcal{H}$	~p
a. ta.qaktk^h	−1		−100	1e-40
☞ b. ta.qa.katk^h		−1	−8.17	1

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

(16) a. *Mid outweighs Ident[high] and/or Max(V).
b. Ident[high] and Ident[high]$_{\sigma1}$ together outweigh *Mid.
c. Max(V) and Max(V)$_{\sigma1}$ together outweigh *Mid.
d. *ɪ outweighs Ident[ATR] and Ident[ATR]$_{\sigma1}$.
e. Phonotac 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 the two sets of URs in (17).[5]

(17) a. *Only concrete URs*
Without abstract URs available, the learner fails to converge on a single UR, and fails to model the data.
b. *All abstract URs*
The learner prefers the more restrictedly abstract UR (/ʔeːw̥e/) to the never surfacing abstract UR (/ʔeːw̥ɪ/).

[5] In the simulations implemented below, the set of possible URs for each morpheme was provided 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 would greatly expand the search space, so the learner might implement the concepts of local lexica from Merchant & Tesar (2008) and 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 Table II, it assigns equal probability to both concrete URs.

		initial (from Table I)	final
Ident[high]	faithfulness	45.69	45.76
Ident[high]$_{\sigma1}$	faithfulness	45.69	45.76
Max(V)	faithfulness	45.31	45.63
Max(V)$_{\sigma1}$	faithfulness	45.31	45.63
*Mid	markedness	85.44	85.60
Ident[ATR]	faithfulness	0.00	0.00
Ident[ATR]$_{\sigma1}$	faithfulness	0.00	0.00
Dep(V)	faithfulness	8.17	8.17
*ɪ	markedness	100.00	100.00
Phonotac	markedness	100.00	100.00

UR	p
/ʔeːw̥i/	0.5
/ʔeːw̥/	0.5

Table II

Constraint weights learned at the second stage, with only concrete URs.

The tableaux in (18) and (19) show the grammar with the constraint weightings and UR probabilities learned in Table II. The selected URs are shown in the leftmost column, together with 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. The probability shown here is the product of the probability of the UR (p(UR)) and the probability of the output given that UR (p(*output*|UR)).

(18) ‘deep’-/ta/

	ʔeːw̥i-Ca p(UR) = 0.5	Max(V) 45.63	Dep(V) 8.17	$\mathcal{H}$	~p
a.	i. ʔeːw̥iCa			0	0.5
	ii. ʔeːw̥Ca	−1		−45.63	1.5e-20
b.	ʔeːw̥-Ca p(UR) = 0.5				
	i. ʔeːw̥Ca			0	0.5
	ii. ʔeːw̥aCa		−1	−8.17	2.8e-4

(18a.i), [ʔeːw̥iCa], is near categorically chosen as the output, given /ʔeːw̥i-Ca/ as the input, as its harmony score is 45.63 better than its competitor

[ʔeːw̥Ca], but the probability of the grammar selecting /ʔeːw̥i-Ca/ → [ʔeːw̥iCa] is only 0.5, because the UR probability is 0.5. If /ʔeːw̥-Ca/ is chosen as the input, then (18b.i), [ʔeːw̥Ca], has the greatest probability. Therefore, the choice of surface form is completely dependent on the choice of UR, with both [ʔew̥Ca] and [ʔeːw̥iCa] receiving near 0.5 probability. This is an incorrect result, as the learner has never seen [ʔeːw̥iCa], only [ʔeːw̥Ca].

(19) 'deep'-/tk^h/

		Phonotac	Max(V)	Dep(V)	$\mathcal{H}$	~p
a.	ʔeːw̥i-tk^h p(UR) = 0.5	100	45.63	8.17		
	i. ʔeːw̥itk^h				0	0.5
	ii. ʔeːw̥tk^h	−1	−1		−145.63	1e-63
b.	ʔeːw̥-tk^h p(UR) = 0.5					
	i. ʔeːw̥tk^h	−1			−100	1e-41
	ii. ʔeːw̥atk^h		−1		−8.17	0.5

A similar result is seen in (19). The grammar outputs (19b.ii), /ʔeːw̥-tk^h/ → [ʔeːw̥atk^h], 50% of the time, even though that form is never seen. The correct surface form, (19a.i), is chosen only half the time. Note here that the two tableaux are contradictory; in order to select (a.i), [ʔeːw̥itk^h], more probability must be assigned to /ʔeːw̥i/, but that UR selects the incorrect output form, [ʔeːw̥iCa], in (18).

If the learner fails to converge on a grammar that accurately models the data, as here, it 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. It settles on the forms in which /e/ has a probability close to 1, with the constraint weights shown in Table III.

As the learner has near categorically learned an abstract UR, it is able to model the data with similar categoricity. The simulation above obtained over 0.9 probability for all surface forms.

The critical change in constraint weighting learned during this stage is the difference between Ident[high] and Max(V). As shown above, the weighting learned by the phonotactic grammar has non-initial /e/'s undecided between deleting or raising, with a near 0.5 probability for each option (when Phonotac doesn't interfere). Now Ident[high] outweighs Max(V) by a margin of 5.20. Thus, /e/ deletes over 98% of the time when the phonotactics allow, as in (20).

		initial (from Table I)	final
Ident[high]	faithfulness	45.69	48.26
Ident[high]$_{\sigma1}$	faithfulness	45.69	43.13
Max(V)	faithfulness	45.31	43.01
Max(V)$_{\sigma1}$	faithfulness	45.31	48.15
*Mid	markedness	85.44	85.40
Ident[ATR]	faithfulness	0.00	0.00
Ident[ATR]$_{\sigma1}$	faithfulness	0.00	0.00
Dep(V)	faithfulness	8.17	10.97
*ɪ	markedness	100.00	100.00
Phonotac	markedness	100.00	100.00

UR	p
/ʔeːw̥i/	2e-8
/ʔeːw̥/	5e-7
/ʔeːw̥e/	0.99
/ʔeːw̥ɪ/	8e-9

Table III

Constraint weights learned at the second stage, with all URs.

(20) 'deep'-/ta/

	ʔeːw̥e-Ca p(UR) = 0.999	*Mid 85.40	Id[high] 48.20	Max(V) 43.00	$\mathcal{H}$	~p
a.	i. ʔeːw̥iCa	−1	−1		−133.60	0.0055
	ii. ʔeːw̥Ca	−1		−1	−128.40	0.9945
	iii. ʔeːw̥eCa	−2			−170.80	4e-19
b.	ʔeːw̥i-Ca p(UR) = 2.4e-8					
	i. ʔeːw̥Ca	−1		−1	−128.40	5e-27
	ii. ʔeːw̥iCa	−1			−85.40	2.4e-8

However, it raises to [i] over 99% of the time when deletion is phonotactically illicit, as in (21).[6]

(21) 'deep'-/tk^h/

	ʔeːw̥e-tk^h p(UR) = 0.9999	Phonotac 100	*Mid 85.40	Id[high] 48.20	Max(V) 43.00	$\mathcal{H}$	~p
a.	i. ʔeːw̥itk^h		−1	−1		−133.60	1
	ii. ʔeːw̥tk^h	−1	−1		−1	−228.40	6e-158
	iii. ʔeːw̥etk^h		−2			−170.80	7e-17
b.	ʔeːw̥i-tk^h p(UR) = 2.4e-8						
	i. ʔeːw̥itk^h		−1			−85.40	2.4e-8

[6] Constraint weights that lead to higher probability of the correct repair for /e/ could be learned if the convergence tolerance on the simulation was more stringent.

Note that these tableaux only show the results for the two most probable URs; since most of the probability is assigned to /ʔeːw̥e/, the probability of any of the other URs being chosen is near-negligible.

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

5 Discussion

The simulation data shows that restrictedly abstract URs are preferred to never surfacing URs. Several factors in the learner lead to this result. To understand those factors, we must understand the distinct constraint weighting conditions that cause each type of UR to model the [i] ~ [∅] alternation. In this section the properties underlying the results will first be explored, and their further implications will then be discussed.

5.1 Why are restrictedly abstract URs preferred?

Since all of the abstract segments considered delete when this is phonotactically licit, Max(V) must be weighted below whatever markedness constraint militates against the abstract segment, as shown in (22a). In order to prevent the segment from raising or strengthening to [i], Ident[high] or Ident[ATR] must outweigh Max(V). Finally as shown in (20) and (21), if the Ident constraint is also weighted below the markedness constraint in (22b), and is outweighed by Max(V) and Phonotac together, as in (22c), [i] will be found when the phonotactics prevent deletion, rather than the faithful [e] or [ɪ].

(22) a. Ident[high] (or Ident[ATR]) outweighs Max(V).
b. *Mid (or *ɪ) outweighs Ident[high] (or Ident[ATR]).
c. Max(V) and Phonotac together outweigh Ident[high] (or Ident[ATR]).

The obvious difference between /e/ and /ɪ/ is that a high weighting of Ident[high] is required to model the basic phonotactics of the language in (16b) above (Ident[high] and Ident[high]$_{\sigma1}$ together outweigh *Mid), whereas a high weighting of Ident[ATR] is not. However, it is not obvious that the weighting condition in (16b) should have any effect on the weight of the general faithfulness constraint – the grammar could have the same predictions just by assigning a lot of weight to the specific faithfulness constraint, and none to the general one. However, 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, if there are more observed violations of the specific constraint than expected violations (or *vice versa*) there must also be at least as 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.[7]

Secondly, the L2 Gaussian prior has a preference for spreading the weight among constraints (and preferring a general constraint used multiple times). Recall that this prior is used to prefer grammars where the constraint weights have moved least from their initial weights (high for markedness, zero for faithfulness). Since the prior for faithfulness constraints is proportional to the sum of the squares of the constraint weights when $c_i = 0$, as in (8), in order to minimise the prior, the learner tries to share the weight between the constraints while also obtaining a weighting condition like that in (16b). Imagine that the sum of the constraints had to reach 80 in order to categorically show [e] surfacing in initial syllables: if all the weight is assigned to 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 (16b) and the two in (23).

(23) a. Ident[high] and Ident[high]$_{\mathrm{V}ː}$ together outweigh *Mid (to protect long [eː]).

b. Ident[high] and Ident[high]$_{\mathrm{N}}$ together outweigh *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 the general constraint, Ident[high], and 20 for each of the specific constraints. This property is generalisable, given a typologically predicted positional privilege pattern – disjunctive licensing as in Klamath, conjunctive licensing where a segment appears *only* as long in the initial syllable of nouns, or any combination of the two – the prior will prefer to assign more weight to 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. /ʔeːw̥e/ is more restrictedly abstract than /ʔeːw̥ɪ/ in Klamath, because while both have segments (/e/ and /ɪ/ respectively) that do not appear in any surface exponent of the morpheme, /e/ is able

[7] This type of argument is more thoroughly explored in recent work by Hughto *et al.* (2015), Pater (2016) and Staubs *et al.* (2016) with respect to an agent-based model of learning, which makes several different assumptions from the MaxLex model, not covered here, but similarly finds a bias towards general rather than specific constraints in a MaxEnt-based framework.

to surface in more positions than /ɪ/. 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 select the grammar and lexicon that prefer more restrictedly abstract forms. The grammar with /ʔeːw̥e/ and the one with /ʔeːw̥ɪ/ as the UR for [ʔeːw̥Ca] ~ [ʔeːw̥itkʰ] perform equally well with respect to the likelihood of the output data. Since both grammars can be described in HG (O'Hara 2015), the probability of the output data given the grammar and lexicon become near categorical regardless of which UR is chosen, and therefore both negative log likelihoods approach zero. Thus the only difference between these two grammar–lexicon pairs is the value which the prior assigns: basically the sum of the weights of the faithfulness constraints squared. In both grammars, since [e] surfaces everywhere but in non-initial syllables, the weighting condition in (16b) must be respected. Ident[high] must therefore have some non-zero positive value in both grammars.

The learner must check the value of the prior for both the grammar that uses /e/ and the one that uses /ɪ/. The phonotactic learner in Table I assigns a weight of 40.13 to Ident[high], to account for the distribution of [e] in the training data. On the other hand, since Ident[ATR] is never respected in the surface data, the constraint can be set to zero. In order to make a particular /e/ or /ɪ/ repair to [i] in contexts where it cannot delete, the respective faithfulness constraint must outweigh Max(V), as shown in (24) (as seen in the simulation results above in (20)).

(24) a. Ident[high] outweighs Max(V) (for /e/ to show [i] ~ [∅] alternation).
b. Ident[ATR] outweighs Max(V) (for /ɪ/ to show [i] ~ [∅] alternation).

Assume without loss of generality that Max(V) is constant at 45.69 in both grammars – it must have a relatively high weight, 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 constraint weights that minimise the prior while also satisfying both the (simplified) universal weighting conditions in (25), necessary to show the proper surface distributions of [e] and [ɪ], and one of the weighting conditions in (24), to allow an abstract UR to serve for the alternation. In order to find that the UR (/e/ or /ɪ/) deletes categorically, the relevant Ident constraint will be assumed to be at least 50.

(25) a. Max(V) = 45.69 b. Ident[high] ≥ 45.31 c. Ident[ATR] ≥ 0

The grammar necessary for /e/ to serve as the UR must weight Ident [high] at 50.00, and can keep Ident[ATR] at 0. For these three constraints, the prior is proportional to $0^2 + 45.69^2 + 50.00^2 = 4587.58$. On the other hand, if /ɪ/ were to serve as the UR, Ident[ATR] must reach 50.00, while the lowest Ident[high] can be is 45.31. Thus the prior would be proportional to $45.31^2 + 40.11^2 + 45.69^2 = 5749.38$. Since the minimum weight of Ident[high] is greater than that of Ident[ATR], the global minimum must put all of its weight into /ʔeːw̥e/. This example can be fully generalised to any set of weighting conditions where Ident[high] must be higher than Ident[ATR] to satisfy phonotactics, including the weights learned in Table I.

5.2 Further implications

Further generalising 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 and never surfacing URs; rather, it 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 non-initial 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 affect 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 also 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 Ident[high] needs to have a relatively high weight in order to protect /e/ from raising to [i]. Thus another emergent bias is found that replicates an analytical preference: learners prefer to minimise 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 to rule them out as possible URs. The same properties that an analyst might look for when picking an abstract UR for an alternation – feature economy, symmetry, minimising lexical gaps – are in fact emergent biases in a MaxEnt learning framework. If a more restrictedly abstract UR is available, the learner will choose 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 possible (Jarosz 2005, Gouskova & Becker 2013). However, unlike Klamath, there does not exist a vowel that appears on the surface in the language that also has a distributional gap in the position where the alternation occurs. If there are no distributional reasons to pick one UR over the others – in Slavic languages, only never surfacing URs are available (of which there will usually be many) – the learner should have no reason to prefer /ɪ/ to /ĩ/ or anything else. I suggest that other last-resort strategies belong here. 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 from data involving true lexical exceptionality.

If a goal of phonological learning is to have a single underlying form for each morpheme, but that goal is not always met, it is important to know how high a priority having a single underlying form is, and in what cases learners turn to last-resort strategies for exceptionality. If these strategies are considered only after the learning of alternations that can be explained with restrictedly abstract URs (like /e/ in Klamath), it suggests that patterns like Klamath 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. Differences between learners leads to different individuals learning different hidden structures for the same data, which may make some different predictions on very low-frequency items, on the treatment of loanwords or on gradient well-formedness judgements.

REFERENCES

Albright, Adam (2002). *The identification of bases in morphological paradigms*. PhD dissertation, University of California, Los Angeles.

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

Allen, Blake & Michael Becker (2015). Learning alternations from surface forms with sub-lexical phonology. Ms, University of British Columbia & Stony Brook University. Available (May 2017) at http://ling.auf.net/lingbuzz/002503.

Baković, Eric (2009). Abstractness and motivation in phonological theory. *Studies in Hispanic and Lusophone Linguistics* **2**. 183–198.

Barker, M. A. R. (1963). *Klamath dictionary*. Berkeley & Los Angeles: University of California Press.

Barker, M. A. R. (1964). *Klamath grammar*. Berkeley & Los Angeles: University of California Press.

Beckman, Jill N. (1998). *Positional faithfulness*. PhD dissertation, University of Massachusetts, Amherst.

Bowers, Dustin (2015). *A system for morphophonological learning and its consequences for language change*. PhD dissertation, University of California Los Angeles.

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

Coetzee, Andries W. & Joe Pater (2011). The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.) *The handbook of phonological theory*. 2nd edn. Malden, Mass. & Oxford: Wiley-Blackwell. 401–431.

Eisenstat, Sarah (2009). *Learning underlying forms with MaxEnt*. MA thesis, Brown University.

Flora, Marie Jo-Ann (1974). *Palauan phonology and morphology*. PhD dissertation, University of California, San Diego.

Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.

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

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

Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.) *Constraints in phonological acquisition*. Cambridge: Cambridge University Press. 158–203.

Hayes, Bruce & Colin Wilson (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: Waverly Press.

Hughto, Coral, Joe Pater & Robert Staubs (2015). Grammatical agent-based modeling of typology. Paper presented at the GLOW Workshop on Computation, Learnability and Phonological Theory, Paris. Slides available (May 2017) at http://blogs.umass.edu/pater/files/2011/10/hughto-pater-staubs-glow.pdf.

Itô, Junko & Armin Mester (1999). The phonological lexicon. In Natsuko Tsujimura (ed.) *The handbook of Japanese linguistics*. Malden, Mass. & Oxford: Blackwell. 62–100.

Jäger, Gerhard (2007). Maximum entropy models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.) *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan*. Stanford: CSLI. 467–479.

Jäger, Gerhard & Anette Rosenbach (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. *BLS* **31**. 181–192.

Jarosz, Gaja (2006). *Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory*. PhD dissertation, Johns Hopkins University.

Jesney, Karen & Anne-Michelle Tessier (2011). Biases in Harmonic Grammar: the road to restrictive learning. *NLLT* **29**. 251–290.

Kenstowicz, Michael & Charles Kisseberth (1977). *Topics in phonological theory*. New York: Academic Press.

Kenstowicz, Michael & Charles Kisseberth (1979). *Generative phonology: description and theory*. New York: Academic Press.

Kiparsky, Paul (1968). How abstract is phonology? In Osama Fujimura (ed.) *Three dimensions of linguistic* theory. Tokyo: Taikusha. 5–56.

Kraft, Dieter (1988). *A software package for sequential quadratic programming*. Cologne: Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt.

Martinet, André (1968). Phonetics and linguistic evolution. In Bertil Malmberg (ed.) *Manual of phonetics*. Amsterdam: North-Holland. 464–487.

Merchant, Nazarré & Bruce Tesar (2008). Learning underlying forms by searching restricted lexical subspaces. *CLS* **41:2**. 33–47.

O'Hara, Charlie (2015). Positionally abstract underlying representations in Klamath. *CLS* **51**. 397–411.

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. In Alejna Brugos, Manuella R. Clark-Cotton & Seungwan Ha (eds.) *Proceedings of the 29th Boston University Conference on Language Development*. Somerville: Cascadilla. 482–492.

Pater, Joe (2016). Learning in typological prediction: grammatical agent-based modeling. Paper presented at the 42nd Annual Meeting of the Berkeley Linguistics Society.

Pater, Joe & Robert Staubs (2013). Feature economy and iterated grammar learning. Paper presented at the 21st Manchester Phonology Meeting.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith (2012). Learning probabilities over underlying representations. In *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology*. Montreal: Association for Computational Linguistics. 62–71. Available (January 2014) at www.aclweb.org/anthology/W12-2308.

Schane, Sanford (1974). How abstract is abstract? In Anthony Bruck, Robert A. Fox & Michael W. La Galy (eds.) *Papers from the parasession on natural phonology*. Chicago: Chicago Linguistic Society. 297–317.

Smolensky, Paul (1996). The initial state and 'Richness of the Base' in Optimality Theory. Ms, Johns Hopkins University. Available as ROA-154 from the Rutgers Optimality Archive.

Stanton, Juliet (2016). Learnability shapes typology: the case of the midpoint pathology. *Lg* **92**. 753–791.

Staubs, Robert, Jennifer Culbertson, Coral Hughto & Joe Pater (2016). Grammar and learning in syntactic and phonological typology. Poster presented at the 90th Annual Meeting of the Linguistic Society of America, Washington, DC.

Staubs, Robert & Joe Pater (2016). Learning serial constraint-based grammars. In John J. McCarthy & Joe Pater (eds.) *Harmonic Grammar and Harmonic Serialism*. London: Equinox. 369–388.

Tesar, Bruce (2014). *Output-driven phonology: theory and learning*. Cambridge: Cambridge University Press.

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