To appear in the Proceedings of the Twelfth Meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology.

Learning probabilities over underlying representations

Joe Pater[*]
pater@linguist.umass.edu

Robert Staubs[*]
rstaubs@linguist.umass.edu

Karen Jesney[†]
jesney@usc.edu

Brian Smith[*]
bwsmith@linguist.umass.edu

[*]Department of Linguistics
University of Massachusetts Amherst
Amherst, MA 01003 USA

[†]Department of Linguistics
University of Southern California
Los Angeles, CA 90089 USA

Abstract

We show that a class of cases that has been previously studied in terms of learning of abstract phonological underlying representations (URs) can be handled by a learner that chooses URs from a contextually conditioned distribution over observed surface representations. We implement such a learner in a Maximum Entropy version of Optimality Theory, in which UR learning is an instance of semi-supervised learning. Our objective function incorporates a term aimed to ensure generalization, independently required for phonotactic learning in Optimality Theory, and does not have a bias for single URs for morphemes. This learner is successful on a test language provided by Tesar (2006) as a challenge for UR learning. We also provide successful results on learning of a toy case modeled on French vowel alternations, which have also been previously analyzed in terms of abstract URs. This case includes lexically conditioned variation, an aspect of the data that cannot be handled by abstract URs, showing that in this respect our approach is more general.

1 Introduction

Phonological underlying representations (URs) introduce structural ambiguity. For example, a morpheme that alternates in voicing, like the one meaning ‘cat’ in Table 1, could have as its underlying representation /bet/ or /bed/, amongst other possibilities. Underlying /bed/ for surface [bet] requires final devoicing, while intervocalic voicing is required for underlying /bet+a/ for [beda] (/-a/ marks the plural). The ambiguity can often be resolved on the basis of further data. For example, if the language includes both voiced and voiceless consonants intervocalically, as in our toy language which also contains [mota], then intervocalic voicing cannot apply across-the-board. The standard phonological analysis, proposed by Jakobson (1948) for similar data from Russian, would thus posit /bed/ as the underlying form for ‘cat’, as in Table 1, along with a phonological grammar that generates final devoicing.

	UR	SR	Meaning
a.	/bed/	[bet]	cat
b.	/bed+a/	[beda]	cats
c.	/mot/	[mot]	dog
d.	/mot+a/	[mota]	dogs

Table 1: Standard URs for final devoicing

An alternating morpheme can also be given a UR that encodes only the fixed aspects of its structure. For example, ‘cat’ could have as its UR /beT/, where /T/ represents an alveolar plosive unspecified for voicing. The grammar would then fill in its voicing specification appropriately in both contexts, adding [−voice] finally, and [+voice] intervocalically. One use of this underspecification is to capture instances of three-way contrast. For example, the language in Table 2 has consonants that alternate in voicing, as in the singular and plural of ‘cat’, as well as consonants that are both fixed voiceless (‘dog’/‘dogs’) and voiced (‘pig’/‘pigs’). Given the URs shown in Table 2, the surface forms are generated if a grammar fills in voicing on underspecified consonants, and does not change specified ones, as in the analysis of Turkish in Inkelas et al. (1997).

	UR	SR	Meaning
a.	/beT/	[bet]	cat
b.	/beT+a/	[beda]	cats
c.	/mot/	[mot]	dog
d.	/mot+a/	[mota]	dogs
e.	/wid/	[wid]	pig
f.	/wid+a/	[wida]	pigs

Table 2: Underspecified URs and ternary contrast

	UR	SR	Meaning
a.	/bet/	[bet]	cat
b.	/bed+a/	[beda]	cats
c.	/mot/	[mot]	dog
d.	/mot+a/	[mota]	dogs
e.	/wid/	[wid]	pig
f.	/wid+a/	[wida]	pigs

Table 3: UR choice and ternary contrast

There are alternatives to this sort of underspecification. For example, the analysis of Turkish in Becker et al. (2011) posits lexically specific intervocalic voicing, applying to some words but not others. Here we pursue the learning consequences of a proposal in Kager (2008), which involves a grammar that chooses different URs across surface contexts. In this example, /bet/ would be chosen when the morpheme occurs word-finally as in [bet], and /bed/ when it occurs prevocalically, as in [beda] (see Table 3 rows a. and b.). This is a kind of *over*-specification in that the meaning 'cat' has two phonological URs. The non-alternating morphemes /mot/ and /wid/ differ in having only a single UR, with voiceless and voiced final consonants respectively, thus yielding the three-way contrast.

Grammars must be able to choose between URs across surface contexts in order to handle phonologically conditioned suppletive allomorphy - i.e. alternation between forms of a morpheme that are not relatable by a phonological derivation even though the contexts in which each occurs is phonologically defined. The alternation between the forms of the indefinite determiner 'a' and 'an' in English is sometimes analyzed as UR choice, since there is no general process in English of [n] insertion or deletion, but the conditioning context is phonological (vowel- *vs.* consonant-initial following word). That grammars have the power to choose URs in this way is uncontroversial; the only controversies concern the proper formalization of UR choice, and whether particular cases involve UR choice or derivation (Nevins, 2011).

Kager's proposal for ternary contrast is unusual in that it uses UR choice for cases that do seem relatively amenable to analysis in terms of derivations from single URs. Phonologists tend to regard a UR choice analysis as more of a last resort, but as far as we know, there exists no explicit proposal for when an analyst, or a learner, should adopt an analysis with multiple URs for a single morpheme, and when a single UR analysis is required.

One worry about a multiple UR analysis is that it could fail to generalize appropriately. If a learner simply memorized which phonological forms of each morpheme appeared in which contexts, it could fail to extract generalizations, such as the restriction against voicing of word-final consonants in our language in Table 1. This is of course a familiar general issue in learning, and it is the focus of our attention here. We consider a learner to have successfully acquired a language if it finds a grammar that generalizes appropriately, irrespective of the extent to which the learner uses a single phonological UR for each meaning.

Presumably, the assumption that multiple UR analyses of alternations are incompatible with generalization is the basis for their traditional last resort status in phonological theory. However, in at least the grammatical framework that we adopt, and probably in many others, it is possible to construct analyses in which alternations are handled by UR choice, and in which generalizations are still captured. A concrete example is provided by the analysis of the final devoicing language illustrated in Tables 4 and 5, and also by each of the results of the learning simulations presented in sections 3 and 4.

Table 4 shows the distribution over URs that our learner, described with references to precedents in the next section, posits for the final devoicing language. The learner's final grammar is using UR choice to get context-appropriate surface forms of 'cat', as can be seen in rows a. and b. The grammar usually picks /bet/ as the UR for 'cat' when it oc-

	UR	SR	Meaning
a.	/bet/ (0.92) /bed/ (0.08)	[bet]	cat
b.	/bed+a/	[beda]	cats
c.	/mot/	[mot]	dog
d.	/mot+a/	[mota]	dogs

Table 4: Learned URs for final devoicing

Constraint	Devoicing	Contrast
CAT→/bed/	3.65	0
CAT→/bet/	0	0
IDENT-VOICE	6.05	43.62
NO-CODA-VOICE	401.41	39.83
INTER-V-VOICE	1.94	39.83

Table 5: Learned weights

curs finally as in [bet], and almost always picks /bed/ when it occurs prevocalically as in [beda]. This analysis diverges even further from standard phonological practice than Kager's ternary contrast analyses, since we have multiple URs where a single UR analysis would not require underspecification or a lexically specific grammar. Furthermore, in this analysis UR choice is probabilistic, as shown visually in Table 4 row a: /bed/ chosen as the UR in word-final position with probability 0.08. Probabilistic UR choice, which also diverges from the analytic norm in phonology, does not have any observable effect here since the URs neutralize to [bet], but we put it to use in the analysis of French in section 4.

These choices of URs and SRs are being made by a probabilistic weighted constraint version of Optimality Theory (OT) (Prince and Smolensky, 2004), described in the next section. The Input is a string of morphemes ('meanings'), and a candidate is a (UR, SR) pair. Throughout this paper, the candidate URs for a morpheme are all and only its forms observed as SRs (given morphologically segmented words). For the current languages, we include as candidate SRs the identity maps from the URs, and the SRs formed by devoicing any final consonant, or voicing any intervocalic one.

There are three types of constraint. UR constraints (Zuraw, 2000; Boersma, 2001) demand a particular UR for a given morpheme, and are violated when a UR differs from the specified one (Boersma and Zuraw's own formalizations differ somewhat). In Table 5, there are two such constraints, CAT→/bed/ and CAT→/bet/. We omit UR constraints for non-alternating morphemes, since their candidate (UR, SR) pairs always have the same UR, and they always satisfy the single UR constraint. Faithfulness constraints demand (UR, SR) fidelity; here we employ only IDENT-VOICE, which requires a match in voicing specification (McCarthy and Prince, 1999). Finally, Output constraints (AKA Markedness constraints) place demands on the SRs. Here we use NO-CODA-VOICE, which penalizes final voicing, and INTER-V-VOICE, which penalizes an intervocalic voiceless consonant.

Table 5 shows the weights for the constraints that were found for the final devoicing language (Devoicing), and for the language with ternary contrast (Contrast); these yield with high probability the (UR, SR) choices for Tables 4 and 3 respectively. The competition between (/bet/, [bet]) and (/bed/, [bet]) as (UR, SR) pairs for 'cat' illustrates the effects of the first three constraints. The two UR constraints obviously differ in their assessments of the two candidates, as does IDENT-VOICE, which prefers the faithful mapping (/bet/, [bet]) over a voicing change in (/bed/, [bet]). For the final devoicing language, the summed weight of IDENT-VOICE and CAT→/bet/ (6.05) is greater than the weight of CAT→/bed/ (3.65), and so the grammar assigns higher probability to (/bet/, [bet]), as shown in Table 4. For the ternary contrast language on the other hand, the UR constraints have zero weight, and so the decision is fully determined by the relatively high weighted IDENT-VOICE, favoring (/bet/, [bet]).

Even though the learner of the final devoicing language has not acquired the single UR of the traditional phonological analysis, it has acquired a contextually conditioned distribution over UR choices that is appropriate for the learning data. There are weights on the UR constraints that would fail to yield this result. For example, if CAT→/bet/ had a sufficiently high weight relative to the other constraints, then the UR would be fixed as /bet/, and there would be no weighting of the remaining constraints that would pick both [beda] as the highest probability candidate for 'cats', and [mota] as the highest probability candidate for 'dogs'.

Anticipating the discussion of learning in the next section, the weight configuration just described can form a local minimum for our learner. In our simulations, it does not fall into this minimum, nor others like it, when weights are initialized at zero.

The effects of the Output constraints are seen in the choice of URs for 'cat' across phonological contexts in both the final devoicing and ternary contrast languages. NO-CODA-VOICE prefers word-final (/bet/, [bet]) over (/bed/, [bed]), and INTER-V-VOICE prefers intervocalic (/bed+a/, [beda]) over (/bet+a/, [beta]). The high weight on IDENT-VOICE in the ternary contrast language results in very low probability for the unfaithful (UR, SR) mappings (/bed/, [bet]) and (/bet+a/, [beda]). The weights for the coda devoicing language are such that a non-negligible proportion of the probability is reserved for unfaithful (/bed/, [bet]).

Since we have in the case of final devoicing an example of a multiple UR analysis for a language with a phonological regularity, we need to ask whether the grammar generalizes appropriately. The answer is yes. Because of the high weight of NO-CODA-VOICE (401.41) and relatively low weight of IDENT-VOICE (6.05), an underlying voiced obstruent will with extremely high probability map to a surface voiceless one in word-final position. In generating final devoicing this grammar produces predictable relationships between morphologically related words. For example, if a learner with this grammar were to see a plural like [maga] and no singular form, it would posit only /mag/ as the UR for the root. Nonetheless, it would predict with probability near 1 that the singular is pronounced [mak].

Given the observed data from the language in Table 4, it would not have been necessary for the learner to construct a grammar that generalizes in this way. For example, the grammar learned for the ternary contrast language also generates the alternation between [bet] and [bed+a], without producing generalized final devoicing. We thus require a learner with a bias for generalization. Our learner, described in the next section, meets this requirement by incorporating an independently motivated preference for high weighted Output constraints, and low weighted Faithfulness. After describing the learner, we go on to provide simulations for somewhat more complex learning problems.

2 The grammar and learning models

In Maximum Entropy or MaxEnt grammar (Goldwater and Johnson, 2003), the probability of an input/output pair (x_i, y_{ij}) is determined by its *harmony*. The harmony H_{ij} of such a pair is the sum of constraint violations $f_c(x_i, y_{ij})$ scaled by the weights of the constraints w_c.

$$H_{ij} = \sum_c w_c f_c(x_i, y_{ij})$$

This definition of harmony is a common property of grammars that use weighted constraints, as in Harmonic Grammar (Smolensky and Legendre, 2006). A MaxEnt grammar maps harmonies to probabilities, where the probability of a particular output for a particular input $p(y_{ij} \mid x_i)$ is proportional to the exponential of its harmony. These exponentials are normalized within an input, yielding probability distributions.

$$\begin{aligned} p(y_{ij} \mid x_i) =& \frac{1}{Z_i} e^{H_{ij}} \\ Z_i =& \sum_{j'} e^{H_{ij'}} \end{aligned}$$

As discussed above, our output candidates are more elaborate than simple surface forms. Instead, inputs are strings of morphemes and candidates are (UR, SR) pairs. A string of input morphemes x_i can map to an SR y_{ij} in potentially many ways—through many possible URs. Each of these (Input, UR, SR) triples potentially incurs distinct constraint violations. The Input/UR pairing is controlled by the UR constraints, while the UR/SR pairing is controlled by Faithfulness. We thus expand our definition of the probability of a mapping from Input to SR to include all options for the URs z_{ijk}.

$$p(y_{ij} \mid x_i) = \sum_k p(y_{ij}, z_{ijk} \mid x_i)$$

The probabilities $p(y_{ij}, z_{ijk} \mid x_i)$ are defined just as for simple input/output probabilities—they simply include a contribution from candidates on URs. This definition encodes an idea that all URs are potentially valid ways of reaching a particular SR, determined only by the relevant violations of constraints, and does not require a single UR to exist for every Input/SR pairing.

The URs z_{ijk} considered for an input x_i are determined by the UR constraints. A UR z_{ijk} is included in the probability calculation for input x_i only if there exists some constraint $x_i \rightarrow z_{ijk}$. These UR constraints, in turn, rely on observed mappings. For every SR y_{ij} corresponding to an input x_i, we include a UR constraint $x_i \rightarrow y_{ij}$. Thus the candidate URs are simply observed surface forms. In the case of a non-alternating form, only one UR constraint will be included and thus only one UR is entertained. In such cases these constraints are always satisfied; we therefore omit them from our analyses without loss of correctness.

This grammatical framework allows a way of viewing the problem of learning as somewhat agnostic with respect to URs. The learner observes some particular distribution over SRs for a particular input morpheme string and can make any consistent choice about the distribution over URs. It is in this respect that our approach diverges most importantly from prior work on learning URs in Optimality Theory-like frameworks. Our model incorporates ideas from Apoussidou (2007), who uses UR constraints for on-line learning of URs in a probabilistic OT framework, and Eisenstat (2009), who uses a log-linear model very similar to ours. Our approach differs, however, in that learning of unique URs is not taken as a goal.

With the above explicit statement of probabilities, the learner's problem is then to minimize the distinction between its predicted Input/SR distribution and the observed probabilities. For the results presented here, we minimize the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the predicted distribution p_w and observed distribution p^*.

$$D(p^* \mid\mid p_w) = \sum_i \sum_j p^*(y_{ij} \mid x_i) \log \frac{p^*(y_{ij} \mid x_i)}{p_w(y_{ij} \mid x_i)}$$

We use an L2 (Gaussian) prior (Tychonoff and Arsenin, 1977) on the weights. Such a prior introduces a pressure for lower weights, which is especially important for categorical learning cases (in which KL minimization reduces to likelihood maximization). These problems contain probabilities at unity, causing weights to scale arbitrarily high without additional restriction. We used a regularization with $\sigma^2 = 10,000$ for all solutions presented in this paper.

$$w^* = \underset{w}{\operatorname{argmin}} D(p^* \mid\mid p_w) + \frac{1}{2\sigma^2} \sum_c w_c^2$$

We also include in our prior a term that maximizes the sum of the weights of Output constraints, and minimizes the sum of the weights of Faithfulness constraints. The objective function remains bounded from above by the L2 prior, and is also bounded from below by a restriction to non-negative weights. This term is adapted from research on phonotactic learning in OT starting with Smolensky (1996); see further references in Jesney and Tessier (2011). It resembles somewhat the R-measure of Prince and Tesar (2004), but unlike the R-measure this added prior is continuous, improving performance in optimization.

$$\lambda \left(\sum_{f \in F} w_f - \sum_{o \in O} w_o \right)$$

In experimentation, we found that this term was necessary to ensure generalization; the L2 prior alone, even with a smaller variance for Faithfulness than Output constraints, was insufficient. It might be possible to create a more refined version of this term that is sensitive to dependencies between constraints, but this version has sufficed for our purposes. The scaling factor λ controls the relative importance of generalization compared to KL minimization. For the solutions presented here, the value of λ was chosen on the basis of repeated optimizations. λ was decreased gradually until a criterion level of performance was reached. For categorical cases, this criterion level was a likelihood of greater than 0.95. For non-categorical cases, criterion was a sum squared error of less than 0.05. The minimization problem presented here was solved using the L-BFGS-B method (Byrd et al., 1995) as implemented in R (R Development Core Team, 2010), and all optimizations were constrained to use non-negative weights, with weights initialized at zero.[1]

[1] Scripts and input files are available at http://blogs.umass.edu/hgr/examples-and-other-resources-for-perceptron-and-solver-r/.

	/re-/	/ra:-/	/ró-/	/rú:-/
/-se/	[rése]	[rá:se]	[róse]	[rú:se]
/-sá/	[resá]	[rasá]	[rósa]	[rú:sa]
/-só:/	[resó:]	[rasó:]	[róso]	[rú:so]

Table 6: Abstract UR analysis of Tesar's language

3 Stress-length interaction

To illustrate some of the challenges of UR learning, Tesar (2006) provides the toy language in Table 6. The table shows the phonological results of combining four initial, perhaps root, morphemes with three final, perhaps suffix, morphemes. The phonologically relevant differences between the vowels are in length, marked with a colon, and stress, marked with an acute accent. The rows and columns are labeled with the URs that Tesar posits; we will discuss their justification shortly.

Stressed vowels can either be short or long, but there is an absolute surface restriction against stressless long vowels. The stress-alternating morphemes that have long allomorphs, 'ra' and 'so', show a predictable alternation in length: long when stressed, short when stressless. There is also a preference for stress on roots. Although the suffixes 'sa' and 'so' attract stress over roots 're' and 'ra', they lose their stress to fixed stress roots 'ru' and 'ro', and there are no fixed stress suffixes.

Tesar's URs represent the contrastive properties of the morphemes. The contrast between vowels that are long when stressed and those that are always short is encoded as an underlying difference in length. The contrast between the suffixes that attract stress and those that don't is similarly encoded as an underlying difference in stress, as is the contrast between roots that alternate in stress and those that don't. The abstract UR is /ra:/, which never surfaces in that shape due to the restriction against unstressed long vowels. The vowel must be long to contrast with /re/, and stressless to contrast with /rú:/.

We adopt Tesar's Output and Faithfulness constraints. STRESS-ROOT demands stress on the root, and STRESS-SUFFIX demands stress on the suffix. Output words are limited to a single stress, so one of these constraints is always violated. NO-LONG-UNSTRESS is violated by a surface long stressless vowel. NO-LONG penalizes all long vowels. The Faithfulness constraint IDENT-STRESS demands a (UR, SR) match in stress, and IDENT-LONG demands (UR, SR) fidelity in length. We include in addition a set of UR constraints that demand forms corresponding to each of the observed SRs, except for those that have only a single SR, whose UR is fixed. Candidate SRs for each UR were all combinations of stress on either the root or suffix (not both), and faithful and shortened long vowels.

UR	SR	*p*	UR	SR	*p*
/ré+se/ /re+se/	[rése]	0.98 0.02	/re+sá/	[resá]	1
/re+só:/	[resó:]	1	/rá:+se/ /ra+se/	[rá:se] [ráse]	0.99 0.01
/ra+sá/	[rasá]	1	/ra+só:/ /rá:+so/	[rasó:] [rá:so]	0.99 0.01
/ró+se/	[róse]	1	/ró+sa/ /ró+sá/ /ró+sá/	[rósa] [rosá]	0.93 0.07 0.01
/ró+so/ /ró+só:/	[róso] [rosó:]	0.99 0.01	/rú:+se/	[rú:se]	1
/rú:+sa/ /rú:+sá/	[rú:sa]	0.93 0.07	/rú:+so/	[rú:so]	1

Table 7: Learned analysis of Tesar's language

The resulting analysis is shown in Table 7, with probabilities rounded to two decimal points. Candidates whose probabilities round to zero are omitted. In all cases a candidate (UR, SR) pair with the correct SR is given highest probability, and is listed in the first row of each cell. Subsequent rows that contain only a UR have the same SR; identical SRs are omitted to aid readability. Given a probabilistic model like a MaxEnt grammar, one cannot define success on a categorical language like this one in terms of granting $p = 1$ to the correct forms, since this will by definition never happen (unless there is only one candidate in a candidate set). Our objective function is stated in terms of maximizing the summed probability of all (UR, SR) pairs that have the correct SR, and an appropriate criterion is therefore to require that the summed probability over full structures be greater for the correct SR than for any other SR. We thus term this simulation successful. We further note that given a MaxEnt grammar that meets this criterion, one can make the probabilities

Constraint	Weight
No-Long-Unstress	26.43
Stress-Root	26.05
Stress-Suffix	23.50
Ident-Stress	7.66
Ident-Long	6.50
‘sa’→/sá/	5.04
‘so’→/só:/	4.96
‘re’→/re/	3.85
‘ra’→/ra/	3.15
‘ra’→/rá:/	0.25
‘so’→/so/	0.02
‘sa’→/sa/	0
‘re’→/ré/	0
No-Long	0

Table 8: Learned weights for Tesar’s language

of the correct forms arbitrarily close to 1 by scaling the weights (multiplying them by some constant).

The constraint weights for the analysis are shown in Table 8. Both of the faithfulness constraints Ident-Stress and Ident-Long have reasonably high weights, which is expected given the observed contrasts in stress and vowel length across morphemes. The highest probability (UR, SR) mappings are in fact always faithful, with alternations arising from different URs being chosen across phonological contexts.

The crucial case for comparison with the abstract UR analysis is the choice between long stressed /rá:/ and short stressless /ra/, shown with underlining in Table 7. When the morpheme ‘ra’ combines with ‘se’, (/rá:+se/, [rá:se]) is preferred to (/ra+se/, [rasé]), partly because it avoids an Ident-Stress violation on the suffix, and also partly because of the greater weight of Stress-Root than Stress-Suffix. On the other hand, when the input is ‘ra’ and ‘sa’, Ident-Stress is no longer at issue since ‘sa’, unlike ‘se’, provides the option of a stressed UR. In this case, the sum of the weights of the constraints preferring short stressless /ra/ in (/ra+sá/, [rasá]) is greater than for those preferring /rá:/ in (/rá:+sa/, [rá:sa]). The fixed stress roots differ from ‘ra’ in not providing the option of a stressless UR, so that a violation of Ident-Stress would be incurred if the suffix were stressed. While the constraint interactions are more complex here, UR choice succeeds in replacing underspecification in a parallel fashion to the simpler case of the ternary voicing contrast discussed in the introduction.

The Output constraints sensitive to vowel length are in the expected configuration given the restriction of long vowels to stressed syllables: unviolated No-Long-Unstress has a relatively high weight (the highest), while the often-violated No-Long, which penalizes all long vowels, has a relatively low weight (the lowest). Ident-Long is sandwiched in between, with the result that an underlying long vowel that surfaces in a stressed syllable will retain its length, while one that surfaces in a stressless syllable will be realized as short, with probabilities approaching 1.

Because of the availability of UR choice, the mapping from an underlying long vowel to a surface short stressless one that high-weighted No-Long-Unstress generates is never observed in Table 7. However, it is the high probability of this mapping given underlying length and surface stresslessness that ensures that the grammar generalizes appropriately. One paradigmatic regularity in this language is that stressless vowels are short, even when they occur in morphemes whose stressed variants have long vowels. To see how this is captured, imagine that a learner with the grammar in Table 8 were presented with a new morpheme ‘su’ in combination with ‘re’, which resulted in SR [resú:]. Given the segmentation [re+sú:], it would then form the UR /sú:/, containing the long stressed vowel of the only alternant that it had seen. The morpheme ‘ru’ also has a single UR, /rú:/, since it is only observed in the learning data as [rú:]. When these are combined as /rú:+sú:/ the resulting SR will be [rú:su], with probability near 1. That is, the grammar generalizes the length alternations, as well as the stress alternations that occur because of the preference for root over suffix stress.

4 Lexically conditioned variation

Here we apply our model to a case of variation, French vowel deletion, which is formalized in terms of candidate SRs having probabilities intermediate between 1 and 0. This case is of particular interest because the probability of deletion varies across

	Word	**UR**	**SR**	***p***
a.	femelle	/fømɛl/	[fømɛl]	1
b.	semestre	/sVmɛstʁ/	[sømɛstʁ]	0.8
			[smɛstʁ]	0.2
c.	semelle	/sVmɛl/	[sømɛl]	0.5
			[smɛl]	0.5
d.	Fnac	/fnak/	[fnak]	1
e.	breton	/bʁøtɔ̃/	[bʁøtɔ̃]	1

Table 9: Underspecified URs for French and data

UR V	**SR**	***p***	**UR V**	**SR**	***p***
Y	s’mestre	0.08	Y	s’melle	0.04
N	s’mestre	0.15	N	s’melle	0.45
Y	semestre	0.77	Y	semelle	0.47
N	semestre	0.01	N	semelle	0.03
Y	f’melle	0.09	N	F[ø]nac	0.07
Y	femelle	0.91	N	Fnac	0.93
Y	breton	1			

Table 10: Learned analysis of French

Constraint	**Weight**
*CCC	467.26
Max	4.93
‘semestre’→/sømɛstʁ/	4.23
‘semelle’→/sømɛl/	2.71
*[ø]	2.58
‘semelle’→/smɛl/	0.10
‘semestre’→/smɛstʁ/	0.03
Dep	0.00

Table 11: Learned weights for French

words, which can be captured in terms of differences in weights of UR constraints.

In French, the mid-vowel [ø] is variably deleted (this vowel is sometimes called ‘schwa’, though it is not an IPA schwa in most varieties). Like one of the toy voicing languages in section 1, French has a ternary contrast, this time in vowel specification. Words either have a non-alternating [ø] (‘femelle’), an alternating [ø] (‘semestre’, ‘semelle’), or no [ø] (‘Fnac’). The ternary contrast has been analyzed by Anderson (1982) as the result of underspecification.[2] As shown in Table 9, a UR with an underspecified vowel (/V/) is able to be deleted, while a UR with a fully specified vowel (/ø/) is not.

The proportions in Table 9 are partially arbitrary, but accurately reflect the relative probabilities in descriptions such as Dell (1973) and in speaker judgments (Racine, 2007). These show that alternating vowels exhibit a range of deletability. Dictionaries also find the two-way distinction between deleting and non-deleting vowels descriptively inadequate, and a number of experimental and corpus studies find a range of deletion rates across words. Near-minimal pairs in which deletion can occur in both words but at different rates, such as ‘semaine’ and ‘semestre’, show that differences in deletion rates cannot be attributed solely to phonological differences, and must be encoded in the the lexicon.

Although [ø]s can be optionally deleted when preceded by a single consonant as in Table 9, [ø] can never be deleted when its deletion would create a three-consonant sequence within a word, as in ‘breton’ [bʁøtɔ̃]. There are also no words with this sort of three-consonant sequence. In addition to learning the differences in the deletion rates of optional [ø]s, the learner must learn the generalization that an [ø] must be present in the ‘breton’ environment. Given a /CCC/ input, we want the grammar to avoid the three-consonant cluster by inserting a vowel.

The phonological conditioning of deletion in real French is far more complex than our simple sketch, but this simplified version is sufficient for present purposes. We use the following constraints. The Output constraints *[ø] and *CCC militate against [ø] and three-consonant sequences in the SR, respectively. The faithfulness constraint Max requires segments in the UR to be present in the SR (‘no deletion’), while Dep requires SR segments to be in the UR (‘no insertion’). As in the previous sections, UR constraints are only included for morphemes with more than one SR. The learning data consisted of the SRs and probabilities from Table 9.

The resulting analysis is shown in Table 10, using the orthographic convention of marking the lack of a vowel with an apostrophe. The presence of

[2] Anderson (1982) argues that underspecification explains the fact that the alternating vowels can both participate in deletion and alternate with [ɛ], while the non-alternating /ø/ can do neither. However, Morin (1988) presents a number of examples of words that participate in [ɛ]-alternation without participating in deletion.

an underlying vowel is indicated with a 'Y' in the UR column, and its lack with an 'N'. The analysis captures the difference between the rates of [ø] in 'semelle' and 'semestre' as a difference in UR selection. The UR with [ø] is more likely for 'semestre' than 'semelle'. The source of this difference can be seen in the constraint weights in Table 11. The difference between the weights of the UR constraint for 'semestre' requiring the vowel and the one that omits it is greater than that for 'semelle'. The phonological generalization that three-consonant sequences are forbidden is captured by the high weight of *CCC relative to DEP, which means that the grammar will add a vowel to a /CCC/ input.

The contrast between the rates of deletion in 'semelle' and 'semestre' illustrates a widespread phenomenon that is unaddressed by most OT approaches to variation and learning, termed lexically conditioned variation (Coetzee and Pater, 2011). That it is handled in at least this toy version of French is a great benefit of this approach. Underspecification, on the other hand, offers no leverage on this problem, since it provides only a distinction between deleting and non-deleting vowels, and not the finer grained distinctions that the data require.

5 Conclusions

It is a generally unresolved issue how a learner decides whether to use one, or more, URs in an analysis of an alternation. Presumably, learners begin by encoding the various phonological realizations of a morpheme. How, and when, do they decide to collapse these into a single UR? The problem is made more difficult because as noted in the introduction, learners need to consider contextually conditioned UR choice, which is required for at least phonologically conditioned suppletive allomorphy. Previous work on UR learning, including Tesar (2006), abstracts from this issue by allowing only single URs. As a reviewer suggests, a Minimum Description Length criterion might create a bias for fewer URs, but this seems not yet to have been implemented.

In the present approach, phonological generalizations can be acquired even when multiple URs are used, as shown in all of our simulations. This means that the issue raised in the last paragraph can be completely sidestepped by never requiring learners to adopt single URs for alternating morphemes. This approach also sidesteps the difficult issues of choosing which parts of each alternant make up the single UR, and when to leave some structure underspecified. With the French simulation, we have further shown that UR choice handles data that escape underspecification. These advantages suggest that the single UR doctrine, in place since Jakobson (1948), is worth reconsidering, especially in frameworks like OT that can formalize contextual choice of URs without loss of generalization.

One direction for further research is in modeling not only choice between allomorphs, but also their discovery in morpheme segmentation, which involves increasing the size of the hypothesized UR constraint set. Our initial explorations show promise, and this could lead to useful applications in natural language processing, in which MaxEnt models are of course already common. Another extension is to other cases of semi-supervised learning. Here we sum over all of the (UR, SR) pairs corresponding to an observed form. Similar summations can be made over other full structures when the learning data are incomplete: over representations such as syllable structures and syntactic trees, and even over derivations. One such extension that we have explored is to learning 'opacity' (Kiparsky, 1973); see Staubs and Pater (2012) for initial results, which do rely on a type of abstract UR. Finally, one might attempt to model learning of paradigmatic generalizations that are probabilistic across the lexicon, as in Turkish voicing (Becker et al., 2011) - see the related MaxEnt results in Hayes et al. (2009) and Moore-Cantwell (2012).

Acknowledgments

We especially thank Diana Apoussidou and David Smith for their collaboration on earlier presentations of this work, and Mark Johnson for extended discussion. Thanks also to Adam Albright, Paul Boersma, Naomi Feldman, Jeff Heinz, John McCarthy, Paul Smolensky, Colin Wilson and three anonymous reviewers for helpful comments. This research was supported by NSF Grant 0813829 to the University of Massachusetts Amherst, by an NSF Graduate Research Fellowship to Robert Staubs, and a SSHRCC doctoral fellowship to Karen Jesney.

References

Stephen Anderson. 1982. The analysis of french shwa: or how to get something for nothing. *Language*, 58:534–573.

Diana Apoussidou. 2007. *The learnability of metrical phonology*. Ph.D. thesis, University of Amsterdam.

Michael Becker, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language*, 87:84–125.

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

Richard Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16:1190–1208.

Andries Coetzee and Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle, and Alan Yu, editors, *The Handbook of Phonological Theory*, pages 401–431. Blackwell, 2nd edition.

François Dell. 1973. *Les règles et les sons. Introduction à la phonologie générative*. Hermann, Paris, 2nd edition.

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

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

Bruce Hayes, Kie Zuraw, Peter Siptar, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85:822–863.

Sharon Inkelas, Cemil Orhan Orgun, and Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of grammar. In *Derivations and Constraints in Phonology*, pages 393–418. Oxford, Clarendon.

Roman Jakobson. 1948. Russian conjugation. *Word*, 4:155–167.

Karen Jesney and Anne-Michelle Tessier. 2011. Biases in Harmonic Grammar: the road to restrictive learning. *Natural Language and Linguistic Theory*, 29:251–290.

René Kager. 2008. Lexical irregularity and the typology of contrast. In Kristin Hanson and Sharon Inkelas, editors, *The Nature of the Word: Studies in Honor of Paul Kiparsky*, pages 397–432. MIT Press.

Paul Kiparsky. 1973. Abstractness, opacity, and global rules. In Osamu Fujimura, editor, *Three Dimensions of Linguistic Theory*, pages 57–86. TEC, Tokyo.

Solomon Kullback and Richard Leibler. 1951. On information and sufficiency. *Annals of Mathematics and Statistics*, pages 22–79.

John McCarthy and Alan Prince. 1999. Faithfulness and identity in prosodic morphology. In René Kager, Harry van der Hulst, and Wim Zonneveld, editors, *The Prosody-Morphology Interface*, pages 218–309. Cambridge University Press.

Claire Moore-Cantwell. 2012. Over- and undergeneralization in derivational morphology. In *NELS Proceedings*.

Yves-Charles Morin. 1988. De l'ajustement du schwa en syllabe fermée dans la phonologie du français. In Hans Basbøll, Yves-Charles Morin, Roland Noske, and Bernard Tranel, editors, *La phonologie du schwa français*, pages 133–189. John Benjamins, Amsterdam.

Andrew Nevins. 2011. Phonologically conditioned allomorph selection. In Colin Ewen, Beth Hume, Marc van Oostendorp, and Keren Rice, editors, *The Companion to Phonology*, pages 2357–2382. Wiley-Blackwell.

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

Alan Prince and Bruce Tesar. 2004. Learning phonotactic distributions. In René Kager, Joe Pater, and Wim Zonneveld, editors, *Fixing Priorities: Constraints in Phonological Acquisition*, pages 245–291. Cambridge University Press.

R Development Core Team. 2010. R: A language and environment for statistical computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

Isabelle Racine. 2007. Effacement du schwa dans des mots lexicaux: constitution d'une base de données et analyse comparative. In *Proceedings of JEL'2007*, pages 125–130. Université de Nantes.

Paul Smolensky and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. MIT Press.

Paul Smolensky. 1996. The initial state and 'Richness of the Base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Johns Hopkins University.

Robert Staubs and Joe Pater. 2012. Learning serial constraint-based grammars. In John J. McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox Press.

Bruce Tesar. 2006. Faithful contrastive features in learning. *Cognitive Science*, 30:863–903.

Andrey Nikolayevich Tychonoff and V. Y. Arsenin. 1977. *Solutions of ill-posed problems*. Winston, New York.

Kie Zuraw. 2000. *Exceptions and regularities in phonology*. Ph.D. thesis, UCLA.