

Learning Distributions over Underlying Representations

Karen Jesney, Joe Pater, and Robert Staubs

University of Massachusetts, Amherst

{kjesney, pater, rstaubs}@linguist.umass.edu

NECPHON 3, October 24th, MIT

Acknowledgements

This research has been done partially in collaboration with Diana Apoussidou and David Smith. It was supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts, Amherst.

0. Overview

Kager (2009) proposes OT “allomorphy” as an alternative account of phenomena dealt with in terms of abstract URs and lexically specific constraints.

We show that such allomorphic analyses can be learned in an approach to UR learning proposed by Apoussidou (2006), reformalized along similar lines as in Eisenstat (NECPhon 2008, 2009).

Two case studies:

1. Learning “abstract URs” in Turkish (Inkelas, Orgun and Zoll 1997) and Paka (Tesar 2006, NECPhon 2008)
2. Learning lexically conditioned variation in French (Coetzee and Pater to appear, Pater 2008)

Our big picture points:

1. A distinction may not need to be drawn between learning allomorphy and “regular” URs, so long as measures are in place to ensure restrictiveness (e.g. Markedness > Faith bias)
2. The need for a learner to search a space of non-surface existing URs may be weaker than sometimes assumed
3. With a model of grammar that yields variation, distributions over URs extend to cases that abstract URs cannot deal with

Outline:

1. Grammatical assumptions, learning goals, how distributions over URs yield “abstract URs”
2. Learning model and results
3. Lexically conditioned variation

1. “Abstract URs” in Turkish and Paka

Many languages show the following type of alternation in the phonological form of morphemes:

- (1) [mat] “bush”
[mada] “bushes”

Assuming the plural morpheme is /-a/, “bush” has two forms:

- (2) [mat] [mad]

A much-discussed hidden structure problem: which is the underlying (lexical) form?

- (3) /mat/ or /mad/?

Our learner's answer is both: the problem becomes one of finding appropriate weights on constraints favoring each one.

Constraints on URs (Boersma 1999, Apoussidou 2007, Eisenstat 2008):

(4) “bush” → /mat/

Assign a score of -1 if the UR is not /mat/

“bush” → /mad/

Assign a score of -1 if the UR is not /mad/

Core assumption:

- The UR constraints state positive demands for observed allomorphs (morpheme segmentation given)
- We'll come back to this in the conclusion (and discussion, presumably)

(5)

“bush”	*CODA- VOICE	IDENT	/mad/	/mat/
/mat/ [mat]			-1	
/mad/ [mat]		-1		-1
/mad/ [mad]	-1			-1

The grammar defines a probability distribution over candidates. Given the data in (1), the learning goal for the tableau in (5) is:

(6) $p (/mat/, [mat]) + p (/mad/, [mat]) = 1$

For instance (for now, highest score gets $p = 1$):

(7)

	"bush"	*CODA-VOICE 1	IDENT 1	/mad/ 1	/mat/ 10	
$p = 1$	/mat/ [mat]			-1		-1
	/mad/ [mat]		-1		-1	-11
	/mad/ [mad]	-1			-1	-11

This solution will not necessarily be appropriate given other data in the language.

For example, if it is a “final devoicing” language like Dutch:

- (8) [mat] “bush”
- [mad+a] “bushes”
- [bat] “tree”
- [pata] “flower”

The problem (given the standard constraint set):

- (9) With a weighting like that in (7), which prefers /mat/ categorically over /mad/, the learner is forced to treat [mad+a] as intervocalic voicing, which is inconsistent with [pata].

A non-standard solution:

- Allow freer selection between the two underlying forms

(10)

	“bush”	*CODA- VOICE 4	IDENT 3	/mad/ 1	/mat/ 1	
$p = 1$	/mat/ [mat]			-1		-1
	/mad/ [mat]		-1		-1	-3
	/mad/ [mad]	-1			-1	-4

(11)

	IDENT 3	INTER[+V] 2	/mad/ 1	/mat/ 1	
/mat+a/ [mata]		-1	-1		-3
/mad+a/ [mata]	-1	-1		-1	-6
/mat+a/ [mada]	-1		-1		-4
$p = 1$ /mad+a/ [mada]				-1	-1

This is a non-standard analysis because a different UR is being chosen depending on phonological context; this type of analysis is usually reserved for cases of suppletion with a partially phonologically predictable element (e.g. a/an).

We judge this non-standard solution viable because with *CODA-VOICE > IDENT, it passes the OT “Richness of the Base” test. In less theory-laden terms:

- (12) In Dutch, final devoicing applies across the board, so this pattern should be independent of the UR selected: we want our Dutch learner to learn that final consonants are predictably voiceless

The test:

- (13) Assume the final state learner is exposed to the new word [bada] “flowers”, seen only in its derived form. It’s given [-a] as an affix, and [bad] with the meaning “flower”. The only observed form is [bad] so it infers /bad/ as the UR.

Does it choose surface [bad] or [bat] for UR /bad/ in isolation?

(14)

	“flower”	*CODA-VOICE 4	IDENT 3	
$p = 1$	/bad/ [bat]		-1	-3
	/bad/ [bad]	-1		-3

To get a final state grammar that is sufficiently restrictive in this respect, the usual approach in OT is to impose a Markedness > Faithfulness bias (cf. Jarosz 2006):

- (15) Initial state ranking (Smolensky 1996)
 Persistent bias (Hayes 2004, Prince & Tesar 2004)

Our learner similarly aims to minimize the weight of Faith constraints (following Jesney & Tessier NECPhon 2007, to appear)

We now turn to a simple case of a multiple-UR analysis of an abstract UR.

“Dutch” becomes Turkish-like when we have a word with a final voiced consonant.

(16) [mat] “bush” [mada] “bushes”
 [pat] “tree” [pata] “trees”
 [bad] “flower” [bada] “flowers”

There are essentially three types of consonant – the contrast is sometimes captured with an “archiphoneme”, that is, with underspecification of voicing (Inkelas, Orgun & Zoll 1997)

(17) Alternating /T/ [0voice]
 Fixed voiceless /t/ [-voice]
 Fixed voiced /d/ [+voice]

This sort of case can be dealt with in terms of the “allomorphic” analysis we have just seen for Dutch (Kager 2009).

IDENT must have sufficient weight to allow contrast in all environments, including word-final position

(18)

“flower”	IDENT	*CODA-VOICE	/bad/
	3	2	1
/bad/ [bat]	-1		
/bad/ [bad]		-1	

$p = 1$

-3
-2

But the markedness constraints can still choose between the URs for /mad/ ~ /mat/, thus yielding the alternation.

For example, in word-final position:

(19)

	“bush”	IDENT	*CODA-VOICE	/mad/	/mat/	
		3	2	1	1	
$p = 1$	/mat/ [mat]			-1		-1
	/mad/ [mat]	-1			-1	-4
	/mad/ [mad]		-1		-1	-3

The Paka language designed by Tesar (2006) as a test of theories of UR learning contains a similar, but more complex example of an abstract UR.

(20)

	/re-/	/ri:-/	/'ro-/	/'ru:-/
/-se/	'rese	'ri:se	'rose	'ru:se
/-'si/	re'si	ri'si	'rosi	'ru:si
/-'so:/	re'so:	ri'so:	'roso	'ru:so

Tesar's "paka" language: stressed syllables preceded by single quote (e.g. 'ro), long vowels indicated by colon (e.g. ri:).

The UR /ri:-/ does not exist in any surface form – it surfaces as *short unstressed* [ri] or *long stressed* ['ri:]. It's required because:

- (21)
- a. It is underlyingly *long* in contrast with /re-/
 - b. It is underlyingly *stressless* in contrast with /'ru:-/

In the allomorphic analysis, the morpheme has two URs, /'ri:/ and /ri/.

When it combines with stressless /se/, stress faithfulness prefers /'ri:/.

(22)

	“{'ri:, ri} +se”	STRESS-FAITH 3	MAIN-LEFT 2	/'ri:/ 1	/ri/ 1	
$p = 1$	/'ri:+se/ ['ri:se]				-1	-1
	/ri+se/ ['rise]	-1		-1		-4
	/ri+se/ [ri'se]	-1	-1	-1		-5

With long stressed /-'so:/, Stress-Faith prefers instead /ri/.

(23)	“{'ri:, ri} + 'so:”	STRESS-FAITH 3	MAIN-LEFT 2	/'ri:/ 1	/ri/ 1	
	/'ri:+'so:/ ['ri:so]	-1			-1	-4
	/ri+'so:/ ['riso]	-2		-1		-4
$p = 1$	/ri+'so:/ [ri'so:]		-1	-1		-3

2. Learning model and results

Call the weighted sum of the violations of a candidate its *harmony*, denoted H :

$$H(x) = \sum_{\ell} w_{\ell} v_{\ell}$$

The probability of a given candidate in a log-linear model is the exponential of H divided by the sum of the exponentials of all candidates.

For candidate x and candidate set S this is:

$$p(x) = \frac{e^{H(x)}}{\sum_{x' \in S} e^{H(x')}}$$

We pause briefly to illustrate the model, showing that we now have real-numbered probability distributions over URs:

(24)	“bush”	*CODA- VOICE 6	IDENT 2	/mad/ 2	/mat/ 1	H	e^H
.73	/mat/ [mat]			-1		-2	.14
.27	/mad/ [mat]		-1		-1	-3	.05
<.001	/mad/ [mad]	-1			-1	-7	< .001

See Johnson (NECPhon 2007) for an overview of the learning of log-linear (max-ent) models, and their history in linguistics and NLP.

We may similarly formalize the probability of an overt form given an input. Index each input form by i and each surface form as j .

Thus we are calculating the probability of an overt form j corresponding to an input i , denoted as y_{ij} , given an input x_i .

We write u_{ij} for the harmony of output y_{ij} being generated from input x_i .

$$p(y_{ij} \mid x_i) = \frac{1}{Z_i} e^{u_{ij}}$$
$$Z_i = \sum_{j'} e^{u_{ij'}}$$

A given overt form y_{ij} may correspond to a number of different possible underlying representations.

We index each of the underlying representations with k and write z_{ijk} for the hidden structure k corresponding to an output j and input i . The probability that a grammar grants to an overt form is thus the sum over all possible full structures:

$$\begin{aligned} p(y_{ij} \mid x_i) &= \sum_k p(y_{ij}, z_{ijk} \mid x_i) \\ &= \frac{1}{Z_i} \sum_k e^{u_{ijk}} \\ Z_i &= \sum_{j'k'} e^{u_{ij'k'}} \end{aligned}$$

In learning, we minimize the difference between the expected probabilities and the target probabilities.

We do this by minimizing the Kullback–Leibler divergence between the distributions.

Calling our model probabilities p_w and the target probabilities p^* , this is:

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

Minimizing K-L divergence variously called Principle of Minimum Discrimination Information, Principle of Minimum Cross-Entropy, or Minxent.

Minimizing K-L divergence for categorical cases (where $p^* = 1$ for one and only one j and letting $0 \log 0 = 0$) reduces to:

$$\min_w D(p^* || p) = \min_w \sum_i -\log p_w(y_{ij} | x_i) = \max_w \sum_i \log p_w(y_{ij} | x_i)$$

As the sum of log probabilities has the same maximum as the product of probabilities, this is just maximum likelihood.

Our model is thus identical to maximum likelihood estimation in the categorical case.

As stated, many solutions could be made incrementally better by increasing weights towards infinity.

To prevent this, we penalize higher weights with a Gaussian regularization term. This biases weights to cluster towards 0.

This makes our objective function the following:

$$\min_{w'} D(p^* || p) + \frac{1}{2\sigma^2} \sum_{\ell} w_{\ell}^2$$

Finally, we want a general Faithfulness < Non-Faithfulness bias.

To do this, we maximize the difference between the sums of these two classes. This is similar to measuring the constraint set's R-measure (Prince and Tesar 2004).

Call the class of Faithfulness constraints F and non-Faithfulness M . Combining this with K-L divergence and regularization, our final objective is:

$$\min_{w'} D(p^* \| p) + \left(\frac{1}{2\sigma^2} \sum_{\ell} w_{\ell}^2 \right) + \left(\sum_{w_f \in F} w_f - \sum_{w_m \in M} w_m \right)$$

The *Paka* simulation had the observed forms shown in (20), and the constraint set in (25).

(25) *Constraints*

MAINSTRESSLEFT	Stress is on the leftmost syllable
MAINSTRESSRIGHT	Stress is on the rightmost syllable
WEIGHTTOSTRESS	Long vowels are stressed
*V:	Vowels are short
IDENTSTRESS	Corresponding input and output vowels have identical stress
IDENTLENGTH	Corresponding input and output vowels have identical length

Learned weights (initial weights 1, variance 46):

(26)	WEIGHTTOSTRESS	63.51
	MAINSTRESSLEFT	58.61
	MAINSTRESSRIGHT	34.10
	IDENTSTRESS	30.98
	IDENTLENGTH	14.99
	*V:	0.26
	/re/ 29.94, /re/ 62.27	/ri:/ 38.80, /ri/ 53.91
	/ro/ 46.35	/ru:/ 46.35 /se/ 46.35
	/si/ 45.66	/si/ 18.11
	/so:/ 96.68	/so/ 0

Our candidate sets included the URs demanded by the UR constraints, and all logically possible SRs with stress placed on one of the syllables, and vowels either short or long.

Mean probability of observed forms/winners:

(27) 0.98

The richness of the base test supplied all possible combinations of stress and vowel length as URs.

With the learned weights, the probability of unattested surface forms (i.e. unstressed long vowels) vanishes.

3. Lexically conditioned variation

We can make use of the fact that this model can encode probabilities over UR choices to handle data that escape abstract URs.

French “schwa” is classic example of a case for which phonologists often posit an abstract segment. A word like *semaine* or *melon* can optionally be pronounced without the first vowel, while the phonetically identical vowel in *belon* cannot undergo deletion. French also has consonant clusters with no intervening vowel, as in *blonde* or *SMIC*.

- (28) *la s(e)maine* ‘the week’
 le m(e)lon ‘the melon’
 la belon ‘the oyster (a particular kind)’
 la blonde ‘the blonde’
 le SMIC ‘the unemployment insurance’

Analysis in terms of abstract UR:

- (29) Alternating vowel: “V” – underspecified vowel
Fixed vowel: /ə/ (or /œ/, /ø/) – fully specified vowel

This analysis is empirically inadequate. Not only is there a distinction between words that alternate and those that don't, but words that alternate differ in terms of how probable the pronunciation without the vowel is

- (30) *le s(e)mestre* low probability of deletion
la s(e)maine high probability of deletion

Weights on URs can be used to encode this sort of fine-grained lexical conditioning on the probability that a phonological process will apply.

There is plenty of evidence for the above empirical claim:

- (31) a. Dictionaries find the two-way categorization of alternating/non-alternating inadequate (Walker 1996)
- b. Corpus-based studies note that some words show greater frequency of deletion than others (e.g. Hansen 1994, Eychenne 2007, Eychenne and Pustka 2007)
- c. Racine and Grosjean (2002) provide data from a production study showing a wide range of deletion frequencies across words
- d. Racine (2007) shows speakers can judge relative deletability

There are also phonological constraints on schwa deletion. Famously, it usually does not create a triconsonantal cluster (Grammont's "loi des trois consonnes" has a you-tube video)

Dell (1973) notes that (p. 207 in 1980 English translation, glosses and transcriptions omitted)

(32) Contrary to what we said above, in very rapid speech the schwa of a small number of words beginning with #C ___ can be dropped even if the preceding word ends in a consonant: *quelle semaine* is sometimes pronounced [kɛlsmɛn]...The other words that have this property in our speech are *je*, *semelle*, *cerise*, *chemise*, *fenêtre* and *petit*...there are other words which always behave according to [the rule above]: *semestre*, *seringue*, *chenille*

Word-internally, this constraint is inviolable in that there are no words like *breton* that ever lose their schwa.

Markedness constraints and candidates:

			*SCHWA	ALIGN	*ADJUNCT
(33)	/V+CəCV/	[V.CəCV]	-1		
		[VC.CV]		-1	
		[V.<C>CV]			-1
/VC+CəCV/	[VC.CəCV]	-1			
	[VC.<C>CV]			-1	
/V+CCəCV/	[V.CCəCV]	-1			
	[VC.<C>CV]		-1	-1	

examples: *la s'maine, quelle s'maine, *le br'ton*

Learning data

(reasonable looking probabilities that can be represented by the grammar)

(34)	<i>probability of deletion</i>
le belon	0
le breton	0
la semaine	.88
quelle semaine	.12
le semestre	.28
quel semestre	0

The learning data contain no syllable structure: the objective function again sums over all phonetically identical structures (e.g. [VC.CV] and [V.<C>CV])

Constraints and learned weights:

(35)	*SCHWA	43.54
	ALIGN	45.75
	*ADJUNCT	48.71
	MAX	0
	/semaine/	44.25
	/s'maine/	47.75
	/semestre/	45.67
	/s'mestre/	46.33

Observed and expected probabilities of deletion given by the final state grammar

(36)	<i>observed</i>	<i>expected</i>
le belon	0	.10
le breton	0	0
la semaine	.88	.80
quelle semaine	.12	.16
le semestre	.28	.25
quel semestre	0	0.02

4. Conclusions

While there is much work remaining to do, we see the prospects for this approach to UR learning as extremely bright.

If this account of these cases of abstract URs goes through, the question is how the account must be elaborated to get other non-trivial URs. Some prospects:

1. URs underlying epenthesis: these may be the by-product of word/morpheme segmentation. The $M > F$ bias may be key here (Alderete, Tesar *et al.*)
2. URs underlying opacity: the “Free Ride” of McCarthy (2005) also involves an $M > F$ bias...

References

- Apoussidou, Diana. 2007. The learnability of metrical phonology. PhD Dissertation. University of Amsterdam.
- Boersma, Paul. 1999. Phonology-semantics interaction in OT, and its acquisition. In Robert Kirchner, Wolf Wikeley & Joe Pater (eds.), *Papers in Experimental and Theoretical Linguistics* 6: 24-35. Edmonton: University of Alberta.
- Coetzee, Andries & Joe Pater. to appear. The place of variation in phonological theory. In John Goldsmith, Jason Riggle & Alan Yu (eds.), *Handbook of Phonological Theory*, 2nd edition. [ROA-946].
- Eisenstat, Sarah. 2008. Learning underlying forms together with constraint weights. Paper presented at NECPhon 2008. Yale University.
- Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, ed. René Kager, Joe Pater & Wim Zonneveld, 158-203. Cambridge: Cambridge University Press. [ROA-327].
- Inkelas, Sharon, C. Orhan Orgun & Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of the grammar. In Iggy Roca (ed.), *Constraints and Derivations in Phonology*, 393-418. Oxford: Clarendon Press.
- Jarosz, Gaja. 2006. *Rich Lexicons and Restrictive Grammars – Maximum Likelihood Learning in Optimality Theory*. PhD dissertation. Johns Hopkins University.

Jesney, Karen & Anne-Michelle Tessier. 2007. Restrictiveness in gradual learning of Harmonic Grammar. Paper presented at NECPhon 2007. University of Massachusetts Amherst.

Jesney, Karen & Anne-Michelle Tessier. to appear. Biases in Harmonic Grammar: the road to restrictive learning. *Natural Language and Linguistic Theory*.

Kager, René. 2009. Lexical irregularity and the typology of contrast. In Kristin Hanson and Sharon Inkelas (eds.), *The Nature of the Word: Essays in Honor of Paul Kiparsky*. Cambridge, MA: MIT Press.

McCarthy, John J. (2005) Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4, 19-56. [Special issue on morphology in phonology, edited by Maria-Rosa Lloret and Jesús Jiménez.]

Pater, Joe. 2008. Lexically-conditioned variation in Harmonic Grammar. Paper presented at OCP-5. Toulouse-Le Mirail.

Prince, Alan & Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing Priorities: Constraints in Phonological Acquisition*, ed. René Kager, Joe Pater & Wim Zonneveld, 245-291. Cambridge: Cambridge University Press. [ROA-353].

Smolensky, Paul. 1996. On the comprehension / production dilemma in child language. *Linguistic Inquiry* 21: 720-731.

Tesar, Bruce. 2006. Faithful contrastive features in learning. *Cognitive Science* 30(5): 863-903.

Tesar, Bruce. 2008. Learning phonological grammars for output-driven maps. Paper presented at NECPhon 2008. Yale University.