

# Learning Hidden Structure with a Log-linear Model of Grammar

Joe Pater   David Smith   Robert Staubs  
Karen Jesney   Ramgopal Mettu

University of Massachusetts, Amherst

January 9, 2010

# Overview

- We propose a method of hidden structure learning using a log-linear model of grammar.
- We apply this to learning over multiple possible underlying representations.
- We achieve good results for learning Tesar's (2006) Paka languages.
- With a simple Markedness  $>$  Faithfulness bias, we obtain generalized patterns which make single abstract underlying representations unnecessary.

## Basic Model

We denote the constraint violations incurred by an input  $x_i$  mapped to an output  $y_{ij}$  by the function  $f(x_i, y_{ij})$ .

The **harmony** of such a mapping is:

$$H_{ij} = \sum_{\ell} w_{\ell} f_{\ell}(x_i, y_{ij}) \quad (1)$$

Thus we utilize an instantiation of Harmonic Grammar (Legendre et al. 1990, Smolensky and Legendre 2006, Pater 2009).

## Probabilities

The probability of a given output candidate  $y_{ij}$  given the input  $x_i$  is the exponential of its harmony normalized by the sum of the exponentials for the corresponding input:

$$p(y_{ij} \mid x_i) = \frac{1}{Z_i} e^{H_{ij}} \quad (2)$$

$$Z_i = \sum_{j'} e^{H_{ij'}} \quad (3)$$

This definition of probability is the same as that adopted in Maximum Entropy grammar (Goldwater and Johnson 2003, Wilson 2006).

## A Paka Example

The Paka languages (Tesar 2006) are a test of theories of UR learning. They consist of roots and suffixes with underlying forms specified for stress and length.

| $x$     | $y$       | $p_y$ | $H$ | FAITH-STRESS<br>1 | MAIN-LEFT<br>3 |
|---------|-----------|-------|-----|-------------------|----------------|
| /re'si/ | [re'si]   | 0.27  | -3  |                   | -1             |
|         | ☞ ['resi] | 0.73  | -2  | -2                |                |

Stressed syllables preceded by single quote (e.g. 'ro).

Long vowels indicated by colon (e.g. ri:).

Throughout, ☞ indicates the overt form of highest probability.

## Multiple Underlying Representations

A widely-held assumption in generative phonology has been that a given lexical item has one and only one underlying/lexical representation.

Say a given suffix  $S$  corresponds to two surface forms  $['si]$  and  $[si]$ . Generally the UR will be analyzed as being one of these.

Instead, a language may have **both** URs for  $S$ , with the grammar selecting between them each time the meaning  $S$  is expressed.

## Constraints on URs

Kager (2009) proposes that OT “allomorphy” might provide an alternative account for phenomena dealt with in terms of single abstract URs or lexically-specific constraints.

Boersma (1999) proposes that there are constraints on the meaning to UR mapping, allowing the grammar to select from among URs.

Apoussidou (2007) elaborates upon this idea and shows how a learner can acquire appropriate rankings of these constraints.

## Formalization

In our formalization, UR constraints have definitions like the following:

$S \rightarrow /si/$ : “the UR of  $S$  is  $/si/$ ”

$S \rightarrow /'si/$ : “the UR of  $S$  is  $/'si/$ ”

These constraints are positively defined.

Only **observed surface forms** have corresponding UR constraints. That is, if  $S$  is never realized as  $[si:]$ , the following constraint is **not** present in constraint set:

$S \rightarrow /si:/$ : “the UR of  $S$  is  $/si:/$ ”

Ongoing work on opacity is investigating enlarging the set of UR constraints in restricted ways (McCarthy 2005).

## Probability Over Hidden Structure

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

The probability assigned to a given overt form  $y_{ij}$  is the sum of the probabilities of all full structures consistent with it:

$$\begin{aligned} p(y_{ij} \mid x_i) &= \sum_k p(y_{ij}, z_{ijk} \mid x_i) & (4) \\ &= \frac{1}{Z_i} \sum_k e^{H_{ijk}} \end{aligned}$$

$$Z_i = \sum_{j'k'} e^{H_{ij'k'}} \quad (5)$$

## Paka with Hidden Structure

| $x$    | $z$     | $y$      | $p_{zy}$ | $p_y$ | ID-S | M-L | /'si/ |
|--------|---------|----------|----------|-------|------|-----|-------|
|        |         |          |          |       | 1    | 3   | 2     |
| {resi} | /re'si/ | [re'si]  | 0.21     | 0.22  |      | -1  |       |
|        | /resi/  |          | 0.01     |       | -1   | -1  | -1    |
|        | /re'si/ | ↖['resi] | 0.57     | 0.78  | -2   |     |       |
|        | /resi/  |          | 0.21     |       | -1   |     | -1    |

ID-S: IDENTSTRESS. Identical stress in UR and output.

M-L: MAINSTRESSLEFT. Left syllable bears main stress.

/'si/: The UR of {si} is /'si/.

## The Learning Objective

We learn the weights  $w^*$  by maximizing the **log likelihood** of the training data.

Denoting the correct output form by  $y_i^*$  and the compatible hidden structures by  $z_{ik}^*$ , we compute the following:

$$w^* = \max_w \sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} \quad (6)$$

Without hidden structure this is just Maximum Entropy learning. A similar approach to UR learning was developed by Eisenstat (2009), who however forced unique URs. Jarosz (2006) develops a distinct approach to learning hidden structure with maximum likelihood.

## Regularization

Unconstrained, weights will tend towards infinity to increase the likelihood of the data (bringing the probability of correct forms closer to 1).

To enforce convergence we introduce an  $L_2$  (Gaussian) regularization term:

$$w^* = \max_w \left[ \sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} - \frac{1}{2\sigma^2} \sum_\ell w_\ell^2 \right] \quad (7)$$

Regularization can prevent optimization from getting trapped in local maxima.

We also establish a hard **minimum** on constraint weights at 0.0 to prevent “beneficial” violations.

## Markedness > Faithfulness Bias

The learner can do well on the objective function by merely memorizing the correct forms—that is, by weighting Faithfulness highly. We thus enforce a simple  $M > F$  bias, following e.g. Smolensky (1996).

We maximize the difference between the sums of the two classes, counting lexical constraints with Markedness. This gives an approximation to Prince and Tesar's (2004) maximization of R-measure.

The factor  $\lambda$  controls the weight of the  $M > F$  term.

## M > F Term

Combined with regularization, this bias keeps the weights of F constraints as low as possible and the weights of M constraints as high possible while maintaining consistency with the target data (cf. Jesney and Tessier, to appear).

For the results presented here, lexical constraints are grouped with M constraints in the set  $\mathcal{M}$ . They are thus also biased to be higher.

$$w^* = \max_w \sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} - \frac{1}{2\sigma^2} \sum_\ell w_\ell^2 \quad (8)$$
$$+ \lambda \left[ \sum_{w_m \in \mathcal{M}} w_m - \sum_{w_f \in \mathcal{F}} w_f \right]$$

## A Paka Case Study

One of Tesar's (2006) Paka languages:

|         | /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  |

## Analyses

In Tesar's analysis there is a UR /ri:-/ which does not exist in any surface form—it surfaces as short unstressed [ri] or long stressed ['ri:]. It's required in his analysis because:

- It is underlyingly long in contrast with /re-/.
- It is underlyingly stressless in contrast with /'ru:-/.

In the allomorphic analysis adopted here, *ri* has two URs: /'ri:/ and /ri/, chosen situationally. The weights of the two UR constraints must be roughly equal. If the difference between the constraints' weights is too large, a trap results in which no distinction can be made between *ri* and either *re* or *ru*.

## Paka 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 |
| /re/, /re:/, ... | Constraints on underlying representations.                  |

## Paka Results

Initial weights at 1.0,  $\sigma^2 = 48.0$ ,  $\lambda = 0.3$ .

|        |         |       |      |           |      |
|--------|---------|-------|------|-----------|------|
| S-LEFT | S-RIGHT | W-S   | *V:  | ID-LENGTH | ID-S |
| 14.88  | 13.92   | 14.40 | 0.00 | 2.56      | 6.26 |

|       |       |       |        |       |        |
|-------|-------|-------|--------|-------|--------|
| /re/  | /'re/ | /ri/  | /'ri:/ | /'ro/ | /'ru:/ |
| 15.15 | 13.65 | 16.51 | 12.29  | 14.40 | 14.40  |

|       |      |       |       |        |
|-------|------|-------|-------|--------|
| /se/  | /si/ | /'si/ | /so/  | /'so:/ |
| 14.40 | 7.41 | 10.99 | 11.61 | 17.19  |


Fully correct, succeeds on a generalization (Richness of the Base) test—not just memorizing.

## A Test of Result Correctness, 1

One case that formerly required an abstract UR:

$ri + se \rightarrow [{}^{\prime}ri:se], *[{}^{\prime}rise], *[ri{}^{\prime}se]$


The UR  $/{}^{\prime}ri:/$  allows stress to be placed without violating ID-S.

| {rise}   | /ri/<br>16.51 | S-LEFT<br>14.88 | S-RIGHT<br>13.92 | / ${}^{\prime}ri:/$<br>12.29 | ID-S<br>6.26 | <i>H</i> |
|--|---------------|-----------------|------------------|------------------------------|--------------|----------|
|  $/{}^{\prime}ri:se/$<br>[ ${}^{\prime}ri:se$ ] | -1            |                 | -1               |                              |              | -30.43   |
| $/rise/$<br>[ ${}^{\prime}rise$ ]  |               |                 | -1               | -1                           | -1           | -32.47   |
| $/rise/$<br>[ri ${}^{\prime}se$ ]  |               | -1              |                  | -1                           | -1           | -33.43   |

## A Test of Result Correctness, 2

$ri + so \rightarrow [ri'so:], *['ri:so], *['riso]$

The UR /ri/ allows stress to be placed without violating ID-S.

| {riso}   | /ri/<br>16.51 | S-LEFT<br>14.88 | S-RIGHT<br>13.92 | /'ri:/<br>12.29 | ID-S<br>6.26 | <i>H</i> |
|--|---------------|-----------------|------------------|-----------------|--------------|----------|
| /'ri:'so:/<br>['ri:so]   | -1            |                 | -1               |                 | -1           | -36.69   |
| /ri'so:/<br>['riso]  |               |                 | -1               | -1              | -2           | -38.73   |
|  /ri'so:/<br>[ri'so:] |               | -1              |                  | -1              |              | -27.17   |

## Richness of the Base

The results pass a RotB test. That is, when all possible combinations of URs are supplied, the surface forms generated follow the patterns of the language learned.

The target language has no unstressed long vowels. Thus the weights learned should preserve this pattern under any combination of URs.

The highest probability these weights give to an unstressed long vowel is  $6.75 \times 10^{-6}$ .

## Paka Typology Test Lexicon

4 suffixes, 4 roots.

Roots:

| r1     | r2     | r3      | r4   |
|--------|--------|---------|------|
| ra 'ra | re re: | ro 'ro: | 'ru: |

Suffixes:

| s1     | s2     | s3      | s4   |
|--------|--------|---------|------|
| sa 'sa | se se: | so 'so: | 'su: |

6 alternating forms with 2 URs each → 12 UR constraints.

## Paka Typology Test Setup

10,000 random grammars created by assigning random weights.

Candidates chosen by these grammars in Harmonic Grammar used as target languages.

Learner settings:

- Initial weights at 1.0
- $\sigma^2 = 48.0$
- $\lambda = 0.3$

## Paka Typology Correctness Results

The candidate with the highest probability is taken as the winner, regardless of its probability.

Percentage of fully correct languages: 98.61%

Percentage of forms generated correctly per language:

Mean = 99.91%

Median = 100.00%

Standard Deviation = 0.79%

## Paka Typology Distribution Analysis

The patterns learned here are categorical. Ideally, then, the probability of the winner will equal 1.00 and all others will equal 0.00. The entropy of such a distribution will be 0.00 bits.

Mean difference between #1 and #2 probability output: 0.89  
(Target: 1.00)

Mean entropy of output probability distribution: 4.61 bits  
(Target: 0.00 bits. Maximum entropy of data: 39.81 bits)

## Conclusion

We see the prospects for this approach to UR learning as extremely bright, particularly in current work on lexically conditioned variation (e.g. French schwa deletion).

This treatment of hidden structure (not just URs) is similarly part of ongoing work on stress learning:

- Learning stress constraints with hidden (foot) structure.
- Learning syllable weight and stress simultaneously.

# Acknowledgments

This research was carried out partially in collaboration with Diana Apoussidou. It was supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts, Amherst.

## 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, and Joe Pater (eds.), *Papers in Experimental and Theoretical Linguistics* 6: 24-35. Edmonton: University of Alberta.
- Eisenstat, Sarah. 2009. Learning underlying forms with MaxEnt. MA thesis. Brown University.
- Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Spenader, J., Eriksson, A., and Dahl, Ö. (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*: 111-120. Stockholm: Stockholm University.

- Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars - maximum likelihood learning in Optimality Theory. PhD dissertation. Johns Hopkins University.
- Jesney, Karen and 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.
- Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic Grammar - a formal multi-level connectionist theory of linguistic wellformedness: an application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*: 884-891. Cambridge, MA: Lawrence Erlbaum.

- McCarthy, John J. 2005. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4: 19-56.
- Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33: 999-1035.
- Prince, Alan and Bruce Tesar. 2004. Learning phonotactic distributions. In René Kager, Joe Pater, and Wim Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*: 245-291. Cambridge: Cambridge University Press.
- Smolensky, Paul. 1996. On the comprehension / production dilemma in child language. *Linguistic Inquiry* 21: 720-731.
- Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press.
- Tesar, Bruce. 2006. Faithful contrastive features in learning. *Cognitive Science* 30(5): 863-903.