

Learning Hidden Metrical Structure with a Log-Linear Model of Grammar

Jason Naradowsky, Joe Pater, David Smith, Robert Staubs

University of Massachusetts Amherst

narad@cs.umass.edu, pater@linguist.umass.edu, dasmith@cs.umass.edu, rstaubs@linguist.umass.edu

Introduction

Log-linear grammar is a probabilistic extension of Optimality Theory (OT; Prince and Smolensky 1993), or more directly, of Harmonic Grammar (see overviews in Smolensky and Legendre 2006, Pater 2009). Also known as Maximum Entropy grammar, it was first proposed for syntax by Johnson (2002), and subsequently applied to phonology by Goldwater and Johnson (2003), Wilson (2006), Jäger (2007), and Hayes, *et al.* (2008), among others.

Likelihood maximization is guaranteed to be convergent for log-linear grammars, but these guarantees do not hold without access to full structures (Riezler 2000)—similar to the concerns about OT Constraint Demotion algorithms (Tesar and Smolensky 2000).

Eisenstat (2008) provides a general model for the learning of hidden structure in the log-linear framework, and shows that it succeeds on cases of learning of phonological underlying representations (URs).

The convergence of such methods is destroyed by local maxima in the search space. It is yet unknown how frequent these maxima are or how well they can be dealt with by existing techniques in unsupervised learning.

In this work we investigate learning a weight-by-position typology. That is, the languages learned assign weight-sensitive stress, treating coda consonants as moraic or non-moraic.

The moraic level of data is not surface-evident and thus not provided to the learner. The weight specification of CVC syllables is thus hidden structure which must be dealt with by the learner.

Probability Model

Call inputs x (word shapes) and possible outputs for each y (stress patterns) with a violation function f and weights w .

The harmony (or score) of each candidate is then the weighted sum of its violations:

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

A candidate's probability is defined as the normalized exponential of harmony:

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

Expanded to include hidden structures z (moracity specifications), the probability of an overt output structure is simply the sum of all consistent full structures:

$$p(y_{ij} | x_i) = \sum_k p(y_{ij}, z_{ijk} | x_i)$$
$$= \frac{1}{Z_i} \sum_k e^{H_{ijk}}$$
$$Z_i = \sum_{j'k'} e^{H_{ij'k'}}$$

Learning Objective

The set of weights w^* for a language is learned by maximizing the log-likelihood of the data.

We include an additional L_2 (log-space Gaussian) prior to ensure finite solutions to the optimization.

$$w^* = \text{MAX}_w \left[\sum_i \log \frac{1}{Z_i} \sum_k e^{H_{ik}^*} - \frac{1}{2\sigma^2} \sum_{\ell} w_{\ell}^2 \right]$$

Constraints

1. **WEIGHT-TO-STRESS** Assign a violation for every unstressed heavy syllable.
2. **WEIGHT-BY-POSITION** Assign a v. for every non-heavy CVC syllable (Hayes 1995).
3. ***CLASH** Assign a v. for every pair of adjacent stressed syllables (Prince 1983, Selkirk 1984).
4. ***LAPSE** Assign a v. for every pair of adjacent stressless syllables (*Ibid.*).
5. ***FINAL-STRESS** Assign a v. for every stressed final syllable (Hyde 2007).
6. **INITIAL-STRESS** Assign a v. for every stressless initial syllable (*Ibid.*).

Example Problem

The type of language learned in this work is demonstrated in the tableaux below.

Only a limited set of the candidates and constraints is shown.

Input	L1	L2	Overt	Full	WBP	WTS	INITIAL-S	*FINAL-S
CVCVC	1		CV'CVC	CV _μ 'CV _μ C _μ			-1	-1
		1	'CVCVC	CV _μ 'CV _μ C _μ	-1		-1	-1
CVCV			CV'CV	CV _μ 'CV _μ		-1		-1
	1	1	'CVCV	CV _μ CV _μ				
CVCV:	1	1	CV'CV:	CV _μ 'CV _μ			-1	-1
			'CVCV:	CV _μ CV _μ		-1		

Experiment: Learning the Typology

Candidates

- Overt forms: Strings of CV, CV', and CVC syllables up to five syllables with all possible stress patterns including at least one stressed syllable.
- Hidden structure: All possible assignments of morae to coda consonants for each overt form
- Violations of the above constraints were assigned automatically.
- Harmonically bounded and tied candidates were removed from the candidate set.

Language Generation and Learning

- 7,200 sets of random weights were assigned to the constraints.
- The languages described by these weights in Harmonic Grammar were used as the target for learning in the model.
- Parameters: Starting weights at 1.00. $\sigma^2 = 48.00$.

Results: Correctness

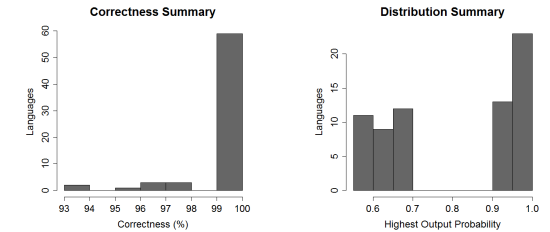
Out of the 7,200 random weightings, 68 unique languages were found.

A grammar is said to be "correct" on an input in a given language if the candidate assigned the highest probability by the grammar matches the form in the target language.

Correctness measures:

1. Languages learned 100% correctly: **59 (86.76%)**
2. Mean correctness percentage: **99.50%**
3. Median correctness percentage: **100.00%**

Most languages were learned well. Those in which some outputs were incorrectly selected deserve further examination.



Results: Distribution

The grammars produced are probabilistic, not categorical. However, the target languages are just the reverse, containing no variation. The target distribution is thus a point distribution—how close do we come?

Three measures:

1. Mean probability of winning candidate: **0.81** (Goal: 1.00)
2. Difference in probability between winner and runner-up: **0.61** (Goal: 1.00)
3. Mean total entropy of output distributions: **174.77 bits** (Goal: 0 bits, MaxEnt: 931.62 bits)

Targets were learned to a high degree of correctness—but the proper distribution was not always learned. This is likely an effect of regularization: the learner is pushed away from categorical solutions by a pressure for low weights.

References

- [1] Eisenstat, S. 2009. Learning underlying forms with MaxEnt. MA thesis. Brown University. [2] Goldwater, S. and M. Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Spender, J., Eriksson, A. and Dahl, O. (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. [3] Hayes, E. 1995. *Metrical Stress Theory: Principles and Case Studies*. [4] Hayes, R., K. Zuraw, P. Siptár, Z. Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). [5] Hyde, B. 2007. The rhythmic foundations of initial Gridmark and Nonfinality. NELS 38. [6] Jäger, G. 2007. Maximum entropy models and Stochastic Optimality Theory. In Zaenen, A., J. Simpson, T. H. King, J. Grimshaw, J. Maling, and G. Manning (eds.), *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan*. [7] Johnson, M. 2002. Optimality-theoretic Lexical Functional Grammar. In Stevenson, S. and P. Merlo (eds.), *The Lexical Basis of Syntactic Processing: Formal, Computational and Experimental Issues*. [8] Pater, J. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33. [9] Prince, A. Relating to the grid. *Linguistic Inquiry* 14(1). [10] Prince, A. and P. Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Ms. Rutgers University & University of Colorado, Boulder. [11] Riezler, Stefan. 2000. Learning log-linear models on constraint-based grammars for disambiguation. In Cussens, J. and Sava Dzerzsovs (eds.), *Learning Language in Logic*. [12] Selkirk, E. O. 1984. *Phonology and Syntax: The Relation Between Sound and Structure*. [13] Smolensky, P. and G. Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. [14] Tesar, B. and P. Smolensky. 2000. *Learnability in Optimality Theory*. [15] Wilson, C. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* 30.5.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0813829. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.