Learning OT Constraint Rankings Using a Maximum Entropy Model

Sharon Goldwater and Mark Johnson
Department of Cognitive and Linguistic Sciences, Brown University
{sharon_goldwater, mark_johnson}@brown.edu

Abstract. A weakness of standard Optimality Theory is its inability to account for grammars with free variation. We describe here the Maximum Entropy model, a general statistical model, and show how it can be applied in a constraint-based linguistic framework to model and learn grammars with free variation, as well as categorical grammars. We report the results of using the MaxEnt model for learning two different grammars: one with variation, and one without. Our results are as good as those of a previous probabilistic version of OT, the Gradual Learning Algorithm (Boersma, 1997), and we argue that our model is more general and mathematically well-motivated.

1. Introduction

One of the requirements of any successful linguistic theory is to provide an explanation of how the learner acquires the language-specific knowledge required by the theory. Optimality Theory (Prince and Smolensky, 1993) is dominant in phonology in part because there are algorithms for learning constraint rankings (Tesar and Smolensky, 1993; Pulleyblank and Turkel, 1996; Prince and Tesar, 1999). Unfortunately, most existing OT learning algorithms have two major problems. First, they are not designed to learn from noisy training data, and generally will not converge when presented with it. Second, because they learn a single OT constraint ranking, they cannot model grammars containing free variation, where a single input form has more than one grammatical output form. (This is a limitation of OT istelf, rather than a weakness of the learning procedures.) In this paper, we concern ourselves with addressing these problems. In particular, we propose that a complete model of phonology and its associated learning algorithm should be able to

- learn from a corpus of real, potentially noisy, data,
- account for free variation as well as categorical distinctions,
- account for effects caused by cumulative constraint violations, and
- generalize to examples not seen in the training data.

There have been various attempts to adapt the OT model in some way to explain free variation, including floating constraints (Nagy and Reynolds, 1997), free ranking of constraints within strata (Anttila, 1997b), and strictness bands (Hayes, 2000). One of the more successful models to date is the probabilistic model proposed by Boersma (1997) and its associated learning algorithm, the Gradual Learning Algorithm. By moving away from the discrete domain of standard OT, the Gradual Learning Algorithm is able to learn from noisy input, and can accurately reproduce grammars with free variation. However, as

This research was supported in part with funding from the National Institute of Mental Health Grant #1R0 IMH60922-01A2.

Keller and Asudeh (2002) have pointed out, the GLA is unable to account for cumulativity effects. Keller's own model, Linear Optimality Theory (Keller, 2000), is designed to account for cumulativity effects, but learns only from acceptability judgment data, not from actual linguistic forms.

In this paper we present a different OT-inspired model of constraint-based phonology, the Maximum Entropy model. This model is in fact a very general statistical model that has been used in many domains and whose mathematical properties are well known. Like the GLA, this model is probabilistic, making it resistant to noise, and seeks to reproduce the distribution of output forms in a training corpus, thus modeling free variation. Like Linear Optimality Theory, the MaxEnt model treats constraints as additive, thus accounting for cumulativity effects.

The connection between OT and Maximum Entropy models used in this paper has been discussed before in Eisner (2000) and Johnson (2002). The estimation procedure or learning method used in this paper is described in detail in Johnson et al. (1999), which also contains statistical consistency results. Johnson (2002) uses the same estimation procedure to learn constraint rankings for OT Lexical Functional Grammars.

The remainder of this paper is organized as follows: We first present the MaxEnt model and its application to constraint-based phonology. We report experimental results similar to those of the GLA on both categorical (no free variation) and stochastic (free variation) training data. We then discuss the question of generalization, explain why it cannot be tested using the kinds of problems presented here, and discuss how we can test for it in future work. Finally, we argue that the MaxEnt model is more general and mathematically simpler than the GLA.

2. The Maximum Entropy Model

Maximum Entropy or log-linear models are a very general class of statistical models that have been applied to problems in a wide range of fields, including computational linguistics. Logistic regression models, exponential models, Boltzmann networks, Harmonic grammars, probabilistic context free grammars, and Hidden Markov Models are all types of Maximum Entropy models. Maximum Entropy models are motivated by information theory: they are designed to include as much information as is known from the data while making no additional assumptions (i.e. they are models that have as high an entropy as possible under the constraint that they match the training data). Suppose we have some conditioning context x and a set of possible outcomes $\mathcal{Y}(x)$ that depend on the context. Then a Maximum Entropy model defines the conditional probability of any particular outcome $y \in \mathcal{Y}(x)$ given the context x as:

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where} \tag{1}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

In these equations, $f_1(y,x) \ldots f_m(y,x)$ are the values of m different features of the pair (y,x), the w_i are parameters (weights) associated with those features, and $Z(x)$ is a normalizing constant obtained by summing over all possible values that y could take on in the sample space $\mathcal{Y}(x)$. In other words, the log probability of y given x is proportional to a linear combination of feature values, $\sum_{i=1}^{m} w_i f_i(y,x)$.

In the MaxEnt models considered here, x is an input phonological form, $\mathcal{Y}(x)$ is the set of candidate output forms (i.e., $\mathcal{Y}$ is the Gen function) and $y \in \mathcal{Y}(x)$ is some particular candidate output form. For an Optimality Theoretic analysis with m constraints $C_1 \cdots C_m$, we use a Maximum Entropy model with m features, and let the features correspond to the constraints.

Thus the feature value $f_i(y, x)$ is the number of violations of constraint C_i incurred by the input/output pair (y, x). We can think of the parameter weights w_i as the ranking values of the constraints.

Note that this Maximum Entropy model of phonology differs from standard Optimality Theory in that constraint weights are additive in log probability. As a result, many violations of lower-ranked constraints may outweigh fewer violations of higher-ranked constraints. This is a property shared by the recent Linear Optimality Theory (Keller, 2000), as well as the earlier theory of Harmonic Grammar (Legendre et al., 1990), on which OT is based.[1] The property of additivity makes the MaxEnt model more powerful and less restrictive than standard OT. When there is sufficient distance between the constraint weights and a finite bound on the number of constraint violations, the MaxEnt model simulates standard OT (see Johnson (2002) for an explicit formula for the weights). The model can therefore account for categorical grammars where a single violation of a highly ranked constraint outweighs any number of violations of lower ranked constraints. However, by assigning closely spaced constraint weights, the MaxEnt model can also produce grammars with variable outputs, or gradient grammaticality effects caused by cumulative constraint violations (Keller, 2000; Keller and Asudeh, 2002). The GLA is able to model grammars with free variation, but, like standard OT, cannot account for these cases of cumulative constraint violations.

Given the generic Maximum Entropy model, we still need to find the correct constraint weights for a given set of training data. We can do this using maximum likelihood estimation on the conditional likelihood (or *pseudo-likelihood*) of the data given the observed outputs:

$$\mathrm{PL}_{\bar{w}}(\bar{y}|\bar{x}) = \prod_{j=1}^{n} \Pr_{\bar{w}}(Y = y_j | x(Y) = x_j) \tag{2}$$

Here, $\bar{y} = y_1 \dots y_n$ are the winning output forms for each of the n training examples in the corpus, and the x_j are the corresponding input forms. So the pseudo-likelihood of the training corpus is simply the product of the conditional probabilities of each output form given its input form. As with ordinary maximum likelihood estimation, we can maximize the pseudo-likelihood function by taking its log and finding the maximum using any standard optimization algorithm. In the experiments below, we used the Conjugate Gradient algorithm (Press et al., 1992).

To prevent overfitting the training data, we introduce a regularizing bias term, or prior, as described in Johnson et al. (1999). The prior for each weight w_i is a Gaussian distribution with mean μ_i and standard deviation σ_i that is multiplied by the psuedo-likelihood in (2). In terms of the log likelihood, the prior term is a quadratic, so our learning algorithm finds the w_i that maximize the following objective function:

$$\log \mathrm{PL}_{\bar{w}}(\bar{y}|\bar{x}) - \sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2} \tag{3}$$

For simplicity, the experiments reported here were conducted using the same prior for each constraint weight, with $\mu_i = 0$ and $\sigma_i = \sigma$. (For possible theoretical implications of this choice, see Section 4.1.) Informally, this prior specifies that zero is the default weight of any constraint (which means the constraint has no effect on the output), so we can vary how closely the model fits the data by varying the standard deviation, σ. Lower values of σ give a more peaked prior distribution and require more data to force the constraint weights away from zero, while higher values give a better fit with less data, but may result in overfitting the data. In particular, multiplying the number of training examples by a factor of r (while

[1] In fact, the Harmony function from Harmonic Grammar is simply $\log \Pr(y|x)$ in (2) (Smolensky and Legendre, 2002).

Constraint	Weight
*RTRHI	33.89
PARSE[RTR]	17.00
GESTURE[CONTOUR]	10.00
PARSE[ATR]	3.53
*ATRLO	0.41

Table 1. Constraint weights learned by MaxEnt model

keeping the empirical distribution fixed) will yield the same result as reducing σ by a factor of $\sqrt{r}$. In other words, if we vary n and σ but hold $n\sigma^2$ constant, the parameter weights learned by the MaxEnt model will be the same.

3. Experimental Results

We ran experiments on two different sets of data, one categorical and one stochastic. Both datasets are available as part of the Praat program (Boersma and Weenink, 2000). In this section, we describe our experimental results and compare them to the results of the GLA on the same datasets, as reported in Boersma (1999) and Boersma and Hayes (2001).

3.1. Learning a Categorical Grammar

For this experiment, we used the Wolof tongue-root grammar described in Boersma (1999), which includes five constraints:

*RTRHI: High vowels must not have a retracted tongue root (rtr).

*ATRLO: Low vowels must not have an advanced tongue root (atr).

PARSE[RTR]: If an input segment is [rtr], it must be realized as [rtr] in the output.

PARSE[ATR]: If an input segment is [atr], it must be realized as [atr] in the output.

GESTURE[CONTOUR]: Do not change from [rtr] to [atr], or vice versa, within a word.

There are 36 input forms provided with this grammar, each of which is paired with a winning output form and three losing candidates. Boersma (1999) reports the results of a sample run of the GLA on this set of data. The algorithm was presented with 10,000 training examples (uniformly distributed among the 36 input forms) with a plasticity of 1.0 and evaluation noise of 2.0,[2] and learned the following ranking:

*RTRHI $\gg$ PARSE[RTR] $\gg$ GESTURE[CONTOUR] $\gg$ PARSE[ATR] $\gg$ *ATRLO

The learned ranking values are sufficiently far apart that the noisy evaluation hardly ever reranks the constraints, giving an error rate below 0.2 percent for all input forms.

We tested the MaxEnt model using various values of $n\sigma^2$, with training data uniformly distributed among the 36 input forms. Like Boersma (1999), we tested the accuracy of the learner on these same 36 input forms. (We discuss ways to test the generalization abilities of the two algorithms in Section 4.3.) In Table 1, we show the constraint weights learned by the MaxEnt model with $n\sigma^2$ at approximately 1,200,000. With these weights, the average error rate over all input forms is 0.07 percent, and the maximum error rate for any input form is 0.19 percent (comparable to the GLA). If we increase $n\sigma^2$, the error rates drop essentially to

[2] See Boersma and Hayes (2001) for a description of the GLA, including an explanation of the plasticity value and evaluation noise.

zero. Note that the constraint weights learned by the MaxEnt model have the same relative ranking as those learned by the GLA and are spaced out at roughly exponential intervals. This sort of exponential pattern of constraint weights is exactly the pattern that, in the limit, gives rise to the strict domination of Optimality Theory (Johnson, 2002).

3.2. Learning a Stochastic Grammar

For this experiment, we used the data on Finnish genitive plurals described in Boersma and Hayes (2001) (henceforth B&H). This set of data was originally collected by Anttila (1997a; 1997b) from a large text corpus.

In Finnish, there are two possible genitive plural endings—a weak ending (usually /-jen/) and a strong ending (usually /-iden/). Some stems allow only one of the two endings (e.g. kameroiden/*kamerojen 'camera', kalojen/*kaloiden 'fish'), while others are acceptable with either ending (e.g. naapurien/naapureiden 'neighbor'). Among the stems that allow both endings, there are differences in the degree to which one ending is preferred over the other, as measured by corpus frequency. Anttila argues that the use of the weak or strong ending is determined entirely by the phonological properties of the stem. He proposes a number of possible constraints in his analysis, of which B&H use 11. Since our aim is to compare the performance of our algorithm to the results in B&H, we use these same 11 constraints:

C_1 (STRESS-TO-WEIGHT): Stressed syllables must be heavy.

C_2 (WEIGHT-TO-STRESS): Heavy syllables must bear stress.

C_3, C_4, C_5 (*Í, *Ó, *Á): No stressed syllables with underlying high/mid/low vowels.[3]

C_6, C_7, C_8 (*Ĭ, *Ŏ, *Ă): No unstressed syllables with underlying high/mid/low vowels.

C_9 (*H.H): No consecutive heavy syllables.

C_{10} (*L.L): No consecutive light syllables.

C_{11} (*LAPSE): No consecutive unstressed syllables.

The data set in B&H contains 5698 tokens, which comprise all genitive plurals of stems ending in light syllables. (Stems ending in heavy syllables require the strong ending and exhibit no variation, so B&H excludes them from the test of stochastic learning.) The tokens are divided into 22 classes depending on the phonological structure of the stem. For each of these classes, the pattern of constraint violations for the winning candidate and the losing candidate is different. Table 2 shows examples of four words from different stem classes and their patterns of constraint violations.

B&H's characterization of the data is misleading, however. Although each of the 22 classes has a different pattern of constraint violations, the GLA does not consider these patterns directly during the learning process. Rather, it learns from the pattern of *differences* between the violations of the winning output and its corresponding losing candidate. Table 3 shows the pattern of differences for each of the stems in Table 2, obtained by subtracting the vector of constraint violations for the winning candidate from that of the losing candidate. Here, we see that from the algorithm's point of view, stems like 'naapuri' and 'ministeri' do not belong to different classes at all. Reanalyzing B&H's classes in this way, it turns out that in fact there are only eight different classes of stems for which distributions must be learned. Since our algorithm, like the GLA, considers only differences in violations between winning and losing candidates, we consider only these eight collapsed classes in reporting our results.

Table 4 compares the results of the GLA and MaxEnt models on this data set. The "Tokens" column shows the number of tokens in each class, and the "% Majority" column

[3] By "underlying vowels", Anttila means vowels in the stem.

Word	Candidates	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
kala	ká.lo.jen	1	1	0	0	0	0	0	1	0	1	1
	ká.loi.den	1	2	0	0	0	0	0	1	1	0	1
naapuri	náa.pu.ri.en	0	1	0	0	0	1	0	0	0	1	2
	náa.pu.rèi.den	0	1	1	0	0	0	0	0	1	0	0
ministeri	mí.nis.te.ri.en	1	2	0	0	0	1	0	0	0	1	3
	mí.nis.te.rèi.den	1	2	1	0	0	0	0	0	1	0	1
maailma	máa.il.mo.jen	0	2	0	0	0	0	0	1	1	0	2
	máa.il.mòi.den	0	2	0	0	1	0	0	0	3	0	0

Table 2. Constraint violation patterns of four of B&H's classes, with example words

Word	Differences in Constraint Violations										
kala	0	1	0	0	0	0	0	0	1	-1	0
naapuri	0	0	1	0	0	-1	0	0	1	-1	-2
ministeri	0	0	1	0	0	-1	0	0	1	-1	-2
maailma	0	0	0	0	1	0	0	-1	2	0	-2

Table 3. Some of B&H's classes are not distinct

Class	Tokens	% Majority	GLA	MaxEnt
1	1097	100	99.5	99.6
2	1000	100	100.0	100.0
3	923	100	100.0	100.0
4	873	70.7	69.5	69.4
5	821	98.4	100	99.8
6	457	99.6	99.4	98.0
7	436	82.1	81.6	80.5
8	91	50.5	58.0	55.3

Table 4. Results of the GLA and MaxEnt on the Stochastic Grammar

shows the percentage of output forms of that class in the training data belonging to the majority output. For example, in class 2, 100% of the output forms belong to the majority output (in this case, /-iden/), whereas in class 4, the outputs are split 70/30 (the more common ending in this case happens to be /-jen/). The "GLA" and "MaxEnt" columns show the percentage of forms produced by these algorithms that match the majority output forms in the training data. The MaxEnt results are for $n\sigma^2$ = 569,800. The GLA results are those reported in B&H, and reflect an average taken over 100 separate runs of the algorithm. During each run, the algorithm was presented with 388,000 training examples. The distribution of input forms in training was according to their empirical frequencies in the corpus, as was the distribution of output forms for each input. The training examples were presented in five groups. The initial plasticity was set to 2.0, but was reduced after each group of examples, to a final value of 0.002. The noise value began at 10.0 for the first group of training examples, and was set to 2.0 for the remaining examples. In their paper, Boersma and Hayes argue that reducing the plasticity corresponds to the child's decreasing ability to learn with age, but give no justification for the change in noise level. In any case, it is not clear how they chose the particular training schedule they report, or whether other training schedules would yield significantly different results. We discuss these points further in Section 4.4.

4. Discussion

In this section, we discuss some of the theoretical implications of our work and the question of generalization. We then compare the results presented for the GLA and MaxEnt model and argue in favor of the MaxEnt model on formal and practical grounds.

4.1. The Initial State

For many applications of the MaxEnt model, the bias term in the objective function is simply a means of preventing overtraining. Here, we can interpret it on a more theoretical level as a learning bias or assumption about the initial state of acquisition. To keep our initial experiments as simple as possible, we used the same prior for each constraint weight, which corresponds to the assumption that all constraints are equally ranked in the absence of data. However, it is widely believed that in fact children's acquisition begins with markedness constraints outranking faithfulness constraints. This situation could easily be modeled by using priors with different means for the markedness and faithfulness constraints, and setting the means for the markedness constraints to some higher value than those for the faithfulness constraints. In the absence of data, markedness would outrank faithfulness, but as data accumulated indicating otherwise, the strength of the data would overcome the prior, and the faithfulness constraints would become more important. Universal rankings could be modeled similarly by adjusting the priors on various constraints to reflect the desired universal ranking.

4.2. The Learning Path

Unlike the GLA and related approaches, our approach cleanly distinguishes the structure of the model (i.e., the MaxEnt exponential form conditional probability distribution (2) and the objective function (3) to be maximized in learning) from the details of the method(s) that can be used to actually maximize that function. This corresponds to the distinction between Marr's *computational level*, which specifies what is to be computed, and Marr's *algorithmic level*, which specifies the algorithms used to carry out that computation (Marr, 1982). This paper's principal claim is that the constraint weights that maximize (3) define a conditional probability distribution (2) that is as accurate as the distributions inferred by the GLA for the cases investigated here.

Any algorithm for maximizing (3) can in principle be used to find the optimal constraint weights. We used the Conjugate Gradient algorithm because it is a well-known efficient general-purpose algorithm that works well on large systems (for other tasks we have used it with thousands of constraint weights and tens of thousands of training items), but there are a number of other algorithms that could be used instead. For example, *iterative scaling algorithms* are specialized for optimizing MaxEnt objective functions (Berger et al., 1996) but should yield the same results as obtained with the Conjugate Gradient algorithm. *Gradient ascent* is a popular but not very efficient optimization algorithm which may produce human-like learning curves, although we have not investigated this here: again, the constraint weights it converges to should be the same as the ones obtained using Conjugate Gradient.[4] We leave for future work the question of which optimization algorithm best models the human learning path.

[4] This discussion ignores the possibility of multiple local maxima. In fact it is possible to show that the log conditional likelihood is concave, so there is only one global maximum (Berger et al., 1996).

4.3. Generalization

In the machine learning community, it is standard practice to evaluate the generalization ability of a learning algorithm by testing on examples not seen in the training data. This is typically done by partitioning the corpus, training on, say, 90% of the data, and testing on the remaining 10%. For small data sets, this process can be repeated using the other nine possible partitions of the corpus to obtain an average test set performance. For very small data sets, the testing portion may consist of only a single data point. Keller and Asudeh (2002) suggest using exactly these methods to evaluate the generalization ability of the GLA, and at first glance, it seems that we should evaluate the MaxEnt learner in this way.

Upon reflection, however, this sort of experiment doesn't make sense for the learning problems we have seen so far. We could set aside 10% of the 5698 Finnish words for testing, but the learning algorithm doesn't see words, it only sees patterns of violations. Since all the words in the corpus fall into only eight classes of violation patterns, the learning algorithm would have already seen many examples of each class during training, and there would be no need to generalize during testing. Alternatively, we could treat the classes themselves as the data points, and perform a leave-one-out regimen. But that would be like providing a child with input that is missing all words with certain phonological characteristics, and expecting the child to be able to produce those words correctly. This is not the normal course of acquisition.

The reason there is no real generalization problem in the tasks we have seen so far is that much of the work has been done before training even begins. The small number of word classes is due to the fact that linguists have chosen a few relevant constraints by which to characterize each word. One of our stated criteria for a successful learning algorithm is the ability to generalize, but we will not be able to test this ability until we start working on more difficult problems. These would be problems with many more constraints, so that the number of possible combinations of constraint violations would be large enough that the algorithm would not see all of the possibilities during training. We are currently working on finding data for a problem of this type in order to truly test the generalization ability of the MaxEnt learner.

4.4. Comparison to the GLA

We believe there are three key features of the GLA that have caused it to become influential. First is its ability to model variation in the adult grammar. Second is the ability to model the initial state (by setting the initial rankings of faithfulness and markedness constraints to different values). Finally, in at least some cases, the GLA seems to mimic the child's learning path (Boersma and Levelt, 1999). We have shown that the MaxEnt algorithm is able to learn both categorical and stochastic grammars as accurately as the GLA. We have not yet run experiments using different priors or different learning algorithms, but we have shown that it would be easy to use these methods to model different assumptions about the initial state and the learning path.

Given the preliminary nature of our results with regard to the actual process of acquisition, why do we believe the MaxEnt model is worth pursuing as an alternative to the GLA? Our argument is twofold. First, the MaxEnt model is mathematically well-motivated, resting on principles of information theory. It has only a single parameter to set—the ratio of σ, the standard deviation of the prior, to the number of training examples (i.e. how closely the model should fit the data). The GLA, in contrast, has at least two parameters—the ratio of the plasticity value to the number of training examples, and the evaluation noise—and potentially many more, if complicated training schedules like the ones in B&H are used. There seems to be no principled way to choose the parameters for a good training schedule, nor do we know how sensitive the results are to that choice, or whether the GLA is guaranteed to converge. In contrast, there is a clear relationship between $n\sigma^2$ and the accuracy of learning in the MaxEnt model, and many optimization algorithms that could be used, including Conjugate Gradient, have proofs of convergence.

The second advantage of the MaxEnt model is its generality. Unlike the GLA, the MaxEnt model is not designed specifically for OT, and in fact has been used in many other fields for a century since its original introduction in statistical physics. The mathematical properties of the model have been well-studied, it has been shown to be useful for learning in a variety of domains, and in general there is a wide literature available (Jelinek, 1997).

5. Conclusions

In this paper we have presented a new way of modeling constraint-based phonology using the statistical framework of the Maximum Entropy model. We have shown that this model, in conjunction with standard optimization algorithms, can learn both categorical and stochastic grammars from a training corpus of input/output pairs. Its performance on these tasks is similar to that of the GLA. We have not yet added any assumptions about the initial state or learning path taken by the MaxEnt model, but we have described how this could easily be done by changing the priors of the model or the optimization algorithm used.

In addition to these empirical facts about the MaxEnt model, we wish to emphasize its strong theoretical foundations. Unlike the GLA, which is a somewhat ad hoc model designed specifically for learning OT constraint rankings, the MaxEnt model is a very general statistical model with an information theoretic justification that has been used successfully for many different types of learning problems. The MaxEnt model also has fewer parameters than the GLA and does not require complicated training schedules. Given our positive results so far and the success of Maximum Entropy models for other types of machine learning tasks, we believe that this model is worth pursuing as a framework for probabilistic constraint-based phonology.

References

Arto Anttila. 1997a. Deriving variation from grammar: a study of finnish genitives. In F. Hinskens, R. van Hout, and L. Wetzels, editors, *Variation, change and phonological theory*, pages 35–68. John Benjamins, Amsterdam. Rutgers Optimality Archive ROA-63.

Arto Anttila. 1997b. *Variation in Finnish phonology and morphology*. Ph.D. thesis, Stanford Univ.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Paul Boersma and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1):45–86.

Paul Boersma and Clara Levelt. 1999. Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of the 30th Child Language Research Forum*.

Paul Boersma and David Weenink. 2000. Praat, a system for doing phonetics by computer. http://www.praat.org.

Paul Boersma. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the Univ. of Amsterdam*, volume 21, pages 43–58.

Paul Boersma. 1999. Optimality-theoretic learning in the praat program. In *Proceedings of the Institute of Phonetic Sciences of the Univ. of Amsterdam*, volume 23, pages 17–35.

Jason Eisner. 2000. Review of Kager: “Optimality Theory”. *Computational Linguistics*, 26(2):286–290.

Bruce Hayes. 2000. Gradient well-formedness in optimality theory. In J. Dekkers, F. van der Leeuw, and J. van de Weijer, editors, *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press, Oxford.

Frederick Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic ‘unification-based’ grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Mark Johnson. 2002. Optimality-theoretic Lexical Functional Grammar. In Paula Merlo and Susan Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, pages 59–74. John Benjamins, Amsterdam, The Netherlands.

Frank Keller and Ash Asudeh. 2002. Probabilistic learning algorithms and optimality theory. *Linguistic Inquiry*, 33(2):225–244.

Frank Keller. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of gramaticality*. Ph.D. thesis, Univ. of Edinburgh.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. Technical Report 90-5, Institute of Cognitive Science, Univ. of Colorado.

David Marr. 1982. *Vision*. W.H. Freeman and Company, New York.

Naomi Nagy and Bill Reynolds. 1997. Optimality theory and variable word-final deletion in faetar. *Language Variation and Change*, 9:37–55.

William Press, Saul Teukolsky, William Vetterling, and Brian Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England, 2 edition.

Alan Prince and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers Univ.

Alan Prince and Bruce Tesar. 1999. Learning phonotactic distributions. Technical Report TR-54, Rutgers Center for Cognitive Science, Rutgers Univ. Rutgers Optimality Archive ROA-353.

Douglas Pulleyblank and William J. Turkel. 1996. Optimality theory and learning algorithms: The representation of recurrent featural asymmetries. In J. Durand and B. Laks, editors, *Current trends in phonology: Models and methods*, pages 653–684. Univ. of Salford.

Paul Smolensky and Géraldine Legendre. 2002. The harmonic mind: From neural computation to optimality-theoretic grammar. Book draft.

Bruce Tesar and Paul Smolensky. 1993. The learnability of optimality theory: An algorithm and some basic complexity results. Ms., Department of Computer Science and Institute of Cognitive Science, Univ. Of Colorado, Boulder. Rutgers Optimality Archive ROA-2.