Richness of the Base and Probabilistic Unsupervised Learning in Optimality Theory

Gaja Jarosz

Department of Cognitive Science
Johns Hopkins University
Baltimore, MD 21218
jarosz@cogsci.jhu.edu

Abstract

This paper proposes an unsupervised learning algorithm for Optimality Theoretic grammars, which learns a complete constraint ranking and a lexicon given only unstructured surface forms and morphological relations. The learning algorithm, which is based on the Expectation-Maximization algorithm, gradually maximizes the likelihood of the observed forms by adjusting the parameters of a probabilistic constraint grammar and a probabilistic lexicon. The paper presents the algorithm's results on three constructed language systems with different types of hidden structure: voicing neutralization, stress, and abstract vowels. In all cases the algorithm learns the correct constraint ranking and lexicon. The paper argues that the algorithm's ability to identify correct, restrictive grammars is due in part to its explicit reliance on the Optimality Theoretic notion of Richness of the Base.

1 Introduction

In Optimality Theory or OT (Prince and Smolensky, 1993) grammars are defined by a set of ranked universal and violable constraints. The function of the grammar is to map underlying or lexical forms to valid surface forms. The task of the learner is to find the correct grammar, or correct ranking of constraints, as well as the set of underlying forms that correspond to overt surface forms given only the surface forms and the set of universal constraints.

The most well known algorithms for learning OT grammars (Tesar, 1995; Tesar and Smolensky, 1995; Boersma, 1997, 1998; Prince and Tesar, 1999; Boersma and Hayes, 2001) are supervised learners and focus on the task of learning the constraint ranking, given training pairs that map underlying forms to surface forms. Recent work has focused on the task of unsupervised learning of OT grammars, where only unstructured surface forms are provided to the learner. Some of this work focuses on grammar learning without training data (Tesar, 1998; Tesar, 1999; Hayes, 2004; Apoussidou and Boersma, 2004). The remainder of this work tackles the problem of learning the ranking and lexicon simultaneously, the problem addressed in the present paper (Tesar et al., 2003; Tesar, 2004; Tesar and Prince, to appear; Merchant and Tesar, to appear). These proposals adopt an algebraic approach wherein learning the lexicon involves iteratively eliminating potential underlying forms by determining that they have become logically impossible, given certain assumptions about the learning problem.[1] In particular, one simplifying assumption of previous work requires that mappings be one-to-one and onto. This assumption prohibits input-output mappings with deletion and insertion as well as

[1] An alternative algorithm is proposed in Escudero (2005), but it has not been tested computationally.

Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006, pages 50–59,
New York City, USA, June 2006. ©2006 Association for Computational Linguistics

constraints that evaluate such mappings. This work represents a leap forward toward the accurate modeling of human language acquisition, but the identification of a general-purpose, unsupervised learner of OT remains an open problem.

In contrast to previous work, this paper proposes a gradual, probabilistic algorithm for unsupervised OT learning based on the Expectation Maximization algorithm (Dempster et al., 1977). Because the algorithm depends on gradually maximizing an objective function, rather than on wholly eliminating logically impossible hypotheses, it is not crucial to prohibit insertion or deletion.

A major challenge posed by unsupervised learning of OT is that of learning *restrictive* grammars that generate only grammatical forms. In previous work, the preference for restrictive grammars is implemented by encoding a bias into the ranking algorithm that favors ranking constraints that prohibit marked structures as high as possible. In contrast, the solution proposed here involves a combination of likelihood maximization and explicit reliance on *Richness of the Base*, an OT principle requiring that the set of potential underlying forms be universal. This combination favors restrictive grammars because grammars that map a "rich" lexicon onto observed forms with high probability are preferred. The proposed model is tested on three constructed language systems, each exemplifying a different type of hidden structure.

2 Learning Probabilistic OT

While the primary task of the grammar is to map underlying forms to overt forms, the grammar's secondary role is that of a filter – ruling out ungrammatical forms no matter what underlying form is fed to the grammar. The role of the grammar as filter follows from the OT principle of Richness of the Base, according to which the set of possible underlying forms is universal (Prince and Smolensky 1993). In other words, the grammar must be restrictive and not over-generate. The requirement that grammars be restrictive complicates the learning problem - it is not sufficient to find a combination of underlying forms and constraint ranking that yields the set of observed surface forms: the constraint ranking must yield only grammatical forms irrespective of the particular lexical items selected for the language.

In classic OT, constraint ranking is categorical and non-probabilistic. In recent years various stochastic versions of OT have been proposed to account for free variation (Boersma and Hayes, 2001), lexically conditioned variation (Anttila, 1997), child language acquisition (Legendre et al., 2002) and the modeling of frequencies associated with these phenomena. In addition to these advantages, probabilistic versions of OT are advantageous from the point of view of learnability. In particular, the Gradual Learning Algorithm for Stochastic OT (Boersma, 1997, 1998; Boersma and Hayes, 2001) is capable of learning in spite of noisy training data and is capable of learning variable grammars in a supervised fashion. In addition, probabilistic versions of OT and variants of OT (Goldwater and Johnson, 2003; Rosenbach and Jaeger, 2003) enable learning of OT via likelihood maximization, for which there exist many established algorithms. Furthermore, as this paper proposes, unsupervised learning of OT using likelihood maximization combined with Richness of the Base provides a natural solution to the grammar-as-filter problem due to the power of probabilistic modeling to use negative evidence implicitly.

The algorithm proposed here relies on a probabilistic extension of OT in which each possible constraint ranking is assigned a probability P(r). Thus, the OT grammar is a probability distribution over constraint rankings rather than a single constraint ranking. This notion of probabilistic OT is similar to - but less restricted than - Stochastic OT, in which the distribution over possible rankings is given by the joint probability over independently normally distributed constraints with fixed, equal variance. The advantage of the present model is computational simplicity, but the proposed learning algorithm does not depend on any particular instantiation of probabilistic OT.

Tables 1 and 2 illustrate the proposed probabilistic version of OT with an abstract example. Table 1 shows the violation marks assigned by three constraints, A, B and C, to five candidate outputs O_1-O_5 for the underlying form, or input /I/. To compute the winner of an optimization, constraints are applied to the candidate set in order according to their rank. Candidates continue to the next constraint if they have the fewest (or tie for fewest) constraint violation marks (indicated by asterisks). In this way the winning or optimal candidate, the candidate that violates the higher-ranked constraints the least, is selected.

		constraints		
input:	/I/	A	B	C
candidates	O_1	*	*	
	O_2	**		*
	O_3		**	
	O_4		*	**
	O_5	*		**

Table 1. OT Candidates and Constraint Violations

The third column of Table 2 identifies the winner under each possible ranking of the three constraints. For example, if the ranking is A >> B >> C, constraint A eliminates all but O_3 and O_4, then constraint B eliminates O_3, designating O_4 as the winner. The remainder of Table 2 illustrates the proposed probabilistic instantiation of OT. The first column shows the probability P(r) that the grammar assigns to each ranking in this example. The probability of each ranking determines the probability with which the winner under that ranking will be selected for the given input. In other words, it defines the conditional probability $P_r(O_k \mid I)$, shown in the fourth column, of the k^{th} output candidate given the input /I/ under the ranking r. The last column shows the total conditional probability for each candidate after summing across rankings. For instance, O_3 is the winner under two of the rankings, and thus its total conditional probability $P(O_3 \mid I)$ is found by summing over the conditional probabilities under each ranking. The total conditional probability $P(O_3 \mid I)$ refers to the probability that underlying form /I/ will surface as O_3, and this probability depends on the grammar.

P(r)	ranking	winner	$P_r(O_k \mid I)$	$P(O_k \mid I)$
0.20	A>>B>>C	O_4	0.2	0.2
0.15	A>>C>>B	O_3	0.15	0.2
0.05	C>>A>>B	O_3	0.05	
0.10	B>>A>>C	O_5	0.1	0.1
0.00	B>>C>>A	O_2	0.0	0.0
0.50	C>>B>>A	O_1	0.5	0.5

Table 2: Probabilistic OT

In addition to the conditional probability assigned by the grammar, this model relies on a probability distribution P(I | M) over possible underlying forms for a given morpheme M. This property of the model implements the standard linguistic proposition that each morpheme has a consistent underlying form across contexts, while the grammar drives allomorphic variation that may result in the morpheme having different surface realizations in different contexts. Rather than identifying a single underlying form for each morpheme, this model represents the underlying form as a distribution over possible underlying forms, and this distribution is constant across contexts. To determine the probability of an underlying form for a morphologically complex word, the product of the morpheme's individual distributions is taken – the probability of an underlying form is taken to be independent of morphological context. For example, suppose that some morpheme M_k has two possible underlying forms, I_1 and I_2, and the two underlying forms are equally likely. This means that the conditional probabilities of both underlying forms are 50%: $P(I_1 \mid M_k) = P(I_2 \mid M_k) = 50\%$.

In sum, the probabilistic model described here consists of a grammar and lexicon, both of which are probabilistic. The task of learning involves selecting the appropriate parameter settings of both the grammar and lexicon simultaneously.

3 Expectation Maximization and Richness of the Base in OT

This section presents the details of the learning algorithm for probabilistic OT. First, in Section 3.1 the objective function and its properties are discussed. Next, Section 3.2 proposes the solution to the grammar-as-filter problem, which involves restricting the search space available to the learning algorithm. Finally, Section 3.3 describes the likelihood maximization algorithm – the input to the algorithm, the initial state, and the form of the solution.

3.1 The Objective Function

The learning algorithm relies on the following objective function:

$$
\begin{aligned}
P_H(O \mid M) &= \prod_k [P_H(O_k \mid M_k)]^{F_k} \\
&= \prod_k [\sum_j P_H(O_k \,\&\, I_{k,j} \mid M_k)]^{F_k} \\
&= \prod_k [\sum_j P_H(O_k \mid I_{k,j}) P_H(I_{k,j} \mid M_k)]^{F_k}
\end{aligned}
$$

The likelihood of the data, or set of overt surface forms, $P_H(O \mid M)$ depends on the parameter settings, the probability distributions over rankings and underlying forms, under the hypothesis H. It is also conditional on M, the set of observed morphemes, which are annotated in the data provided to the algorithm. M is constant, however, and does not differ between hypotheses for the same data set. Under this model each unique surface form O_k is treated independently, and the likelihood of the data is simply the product of the probability of each surface form, raised to the power corresponding to its observed frequency F_k. Each surface form O_k is composed of a set of morphemes M_k, and each of these morphemes has a set of underlying forms $I_{k,j}$. The probability of each surface form $P_H(O_k \mid M_k)$ is found by summing the joint distribution $P_H(O_k \& I_{k,j} \mid M_k)$ over all possible underlying forms $I_{k,J}$ for morphemes M_k that compose O_k. Finally, the joint probability is simply the product of the conditional probability $P_H(O_k \mid I_{k,j})$ and lexical probability $P_H(I_{K,j} \mid M_k)$, both of which were defined in the previous section.

The primary property of this objective function is that it is maximal only when the hypothesis generates the observed data with high probability. In other words, the grammar must map the selected lexicon onto observed surface forms without wasting probability mass on unobserved forms. Because there are two parameters in the model, this can be accomplished by adjusting the ranking distributions or by adjusting lexicon distributions. The probability model itself does not specify whether the grammar or the lexicon should be adjusted in order to maximize the objective function. In other words, the objective function is indifferent to whether the restrictions observed in the language are accounted for by having a restrictive grammar or by selecting a restrictive lexicon. As discussed in Section 2, according to Richness of the Base, only the first option is available in OT: the grammar must be restrictive and must neutralize noncontrastive distinctions in the language. The next subsection addresses the proposed solution – a restriction of the search procedure that favors maximizing probability by restricting the grammar rather than the lexicon.

3.2 Richness of the Base

Although the notion of a restrictive grammar is intuitively clear, it is difficult to implement formally. Previous work on OT learnability (Tesar, 1995; Tesar and Smolensky, 1995; Smolensky 1996; Tesar, 1998, Tesar, 1999; Tesar et al., 2003; Tesar and Prince, to appear; Hayes, 2004) has proposed the heuristic of *Markedness over Faithfulness* during learning to favor restrictive grammars. In OT there are two basic types of constraints, markedness constraints, which penalize dispreferred surface structures, and faithfulness constraints, which penalize nonidentical mappings from underlying to surface forms. In general, a restrictive grammar will have markedness constraints ranked high, because these constraints will restrict the type of surface forms that are allowed in a language. On the other hand, if faithfulness constraints are ranked high, all the distinctions introduced into the lexicon will surface. Thus, a heuristic preferring markedness constraints to rank high whenever possible does in general prefer restrictive grammars. However, the markedness over faithfulness heuristic does not exhaust the notion of restrictiveness. In particular, markedness over faithfulness does not favor grammar restrictiveness that follows from particular rankings between markedness constraints or between faithfulness constraints.

This work aims to provide a general solution that does not require distinguishing various types of constraints – the proposed solution implements Richness of the Base explicitly in the initial state of the lexicon. Specifically, the solution involves requiring that initial distributions over the lexicon be uniform, or rich. Although the objective function alone does not prefer restrictive grammars over restrictive lexicons, a lexicon constrained to be uniform, or nonrestrictive, will in turn force the grammar to be restrictive. Another way to think about it is that a restrictive grammar is one that compresses the input distributions maximally by mapping as much of the lexicon onto observed surface forms as possible. By requiring the lexicon to be rich the proposed solution relies on the objective function's natural preference for grammars that maximally compress the lexicon. The objective function prefers restrictive grammars in this situation because restrictive grammars will allow the highest probability to be assigned to observed forms. In contrast, if the lexicon is not rich, there is nothing for the grammar to compress, and the objective function's natural preference for compression will not be employed. The next subsection discusses the algorithm and the initialization of the parameters in more detail.

3.3 Likelihood Maximization Algorithm

As discussed above, the goal of the learning algorithm is to find the probability distributions over rankings and lexicons that maximize the probability assigned to the observed set of data according to the objective function. In addition, any regularities present in the data should be accommodated by the grammar rather than by restricting the lexicon. As in previous work on unsupervised learning of OT, the algorithm assumes knowledge of OT constraints, the possible underlying forms of overt forms, and sets of candidate outputs and their constraint violation profiles for all possible underlying forms. While the present version of the algorithm receives this information as input, recent work in computational OT (Riggle, 2004; Eisner, 2000) suggests that this information is formally derivable from the constraints and overt surface forms and can be generated automatically.

In addition, the algorithm receives information about the morphological relations between observed surface forms. Specifically, output forms are segmented into morphemes, and the morphemes are indexed by a unique identifier. This information, which has also been assumed in previous work, cannot be derived directly from the constraints and observed forms but is a necessary component of a model that refers to underlying forms of morphemes. The present work assumes this information is available to the learner although Section 5 will discuss the possibility of learning these morphological relations in conjunction with the learning of phonology.

The set of potential underlying forms is derived from observed surface forms, morphological relations, and the constraint set. On the one hand the set of potential underlying forms, which is initially uniformly distributed, should be rich enough to constitute a rich base for the reasons discussed earlier. On the other hand, the set should be restricted enough so that the search space is not too large and so that the grammar is not pressured to favor mapping underlying forms to completely unrelated surface forms. For this reason, potential underlying forms are derived from surface forms by considering all featural variants of surface forms for features that are evaluated by the grammar. Of these potential underlying forms, only those that can yield each of the observed surface allomorphs of the morpheme under some ranking of the constraints are included. This formulation differs substantially from previous work, which aimed to construct the lexicon via discrete steps, the first of which involved permanently setting the values for features that do not alternate. In contrast, the approach taken here aims to create a rich initial lexicon, to compel the selection of a restrictive grammar.

In addition to featural variants, variants of surface forms that differ in length are included if they are supported by allomorphic alternation. In particular, featural variants of all the observed surface allomorphs of the morpheme are considered as potential underlying forms for the morpheme if each of the observed surface forms can be generated under some ranking. Including these types of underlying forms extends previous work, which did not allow segmental insertion or deletion or constraints that evaluate these unfaithful mappings, such as MAX and DEP.

The algorithm initializes both the lexicon and grammar to uniform probability distributions. This means that all rankings are initially equally likely. Likewise, all potential underlying forms for a morpheme are initially equally likely. Thus, the probability distributions begin unbiased, but choosing an unbiased lexicon initially begins the search through parameter space at a position that favors restrictive grammars. The experiments in the following section suggest that this choice of initialization correctly selects a restrictive final grammar.

The learning algorithm itself is based on the Expectation Maximization algorithm (Dempster et al., 1977) and alternates between an expectation stage and a maximization stage. During the expectation stage the algorithm computes the likelihood of the observed surface forms under the current hypothesis. During the maximization stage the algorithm adjusts the grammar and lexicon distributions in order to increase the likelihood of the data. The probability distribution over rankings is adjusted according to the following re-estimation formula:

$$P_{H+1}(r)=\sum_{k}\frac{F_k}{\sum_{k}F_k}\cdot\frac{P_H(O_k\,|\,r,M_k)}{P_H(O_k\,|\,M_k)}$$

Intuitively, this formula re-estimates the probability of a ranking for state *H+1* in proportion to the ranking's contribution to the overall probability at state *H*. The algorithm re-estimates the probability distribution for an underlying form according to an analogous formula:

$$P_{H+1}(I_{k,j}\,|\,M_i)=\sum_{k}\frac{F_k}{\sum_{k}F_k}\cdot\frac{P_H(O_k\,\&\,I_{k,j}\,|\,M_i)}{P_H(O_k\,|\,M_i)}$$

Intuitively, the re-estimate of the probability of an underlying form $I_{k,j}$ for state *H+1* is proportional to the contribution that underlying form makes to the total probability due to morpheme M_i at state *H*. The algorithm continues to alternate between the two stages until the distributions converge, or until the change between one stage and the next reaches some predetermined minimum. At this point the resulting distributions are taken to correspond to the learned grammar and lexicon.

4 Experiments

This section describes the results of experiments with three artificial language systems with different types of hidden structure. In all experiments presented here, each unique surface form is assumed to occur with frequency 1.

4.1 Voicing Neutralization

The first test set is an artificial language system (Tesar and Prince, to appear) exhibiting voicing neutralization. The constraint set includes five constraints:

- NOVOI - No voiced obstruents
- NOSFV- No syllable-final voiced obstruents
- IVV - No intervocalic voiceless consonants
- IDVOI - Surface voicing must match underlying voicing
- MAX - Input segments must have output correspondents

These five constraints can describe a number of languages, but of particular interest are languages in which voicing contrasts are neutralized in one or more positions. Such languages, three of which are shown below, test the algorithm's ability to identify correct and restrictive grammars. The partial rankings shown below correspond to the necessary rankings that must hold for these languages; each partial ranking actually corresponds to several total rankings of the constraints. Also shown below are the morphologically analyzed surface forms for each language that are provided as input to the algorithm. The subscripts in these forms indicate morpheme identities, while the hyphens segment the words into separate morphemes. For example, tat$_{1,2}$ means that the surface form "tat" could be derived from either morpheme 1 or 2 in this language.

- (A) Final devoicing, contrast intervocalically:
  - NOSFV, MAX >> IDVOI >> IVV, NOVOI
  - tat$_{1,2}$; dat$_{3,4}$; tat$_1$-e$_5$; tad$_2$-e$_5$; dat$_3$-e$_5$; dad$_4$-e$_5$
- (B) Final devoicing and intervocalic voicing:
  - NOSFV, MAX, IVV >> IDVOI, NOVOI
  - tat$_{1,2}$; dat$_{3,4}$; tad$_{1,2}$-e$_5$; dad$_{3,4}$-e$_5$
- (C) No voiced obstruents:
  - MAX, NOVOI >> IDVOI, IVV
  - tat$_{1,2,3,4}$; tat$_{1,2,3,4}$-e$_5$

In language C, it would be possible to maximize the objective function by selecting a restrictive lexicon rather than a restrictive grammar. In particular, /tat/ could be selected as the underlying form for morphemes 1-4 in order to account for the lack of voiced obstruents in the observed surface forms. In this case, the objective function could just as well be satisfied by an identity grammar mapping underlying /tat/ to surface "tat". However, as discussed in Section 2, such a grammar would violate the principle of Richness of the Base by putting the restriction against voiced obstruents into the lexicon rather than the grammar. Thus, this language tests not only whether the algorithm finds a maximum, but also whether the maximum corresponds to a restrictive grammar.

In fact, for all three languages above, the algorithm converges on the correct, restrictive grammars and correct lexicons. Specifically, the final grammars for each of the languages above converge on probability distributions that distribute the probability mass equally among the total rankings consistent with the partial orders above. For example, for language C the algorithm converges on a distribution that assigns equal probability to the 20 total rankings consistent with the partial order given by MAX, NOVOI >> IDVOI, IVV.

The initial uniform lexicon for language C is shown in Table 3. Here the numbers 1-5 refer to morpheme indices, and the possible underlying forms for each morpheme are uniformly distributed. This initial lexicon favors a grammar that can map as much of the rich lexicon as possible onto surface forms with no voiced obstruents. With these constraints, this translates into ranking NOVOI above IDVOI and IVV. As the algorithm begins learning the lexicon and continues to refine its hypothesis for this language, nothing drives the algorithm to abandon the initial rich lexicon. Thus, in the final state, the lexicon for this language is identical to the initial lexicon. In general, the final lexicon will be uniformly distributed over underlying forms that differ in noncontrastive features.

1	/tat/ - 25%	/tad/ - 25%	/dat/ - 25%	/dad/ - 25%
2	/tat/ - 25%	/tad/ - 25%	/dat/ - 25%	/dad/ - 25%
3	/tat/ - 25%	/tad/ - 25%	/dat/ - 25%	/dad/ - 25%
4	/tat/ - 25%	/tad/ - 25%	/dat/ - 25%	/dad/ - 25%
5	/e/ - 100%			

Table 3. Initial Lexicon for Language C

4.2 Grammatical and Lexical Stress

The next set of languages from the PAKA system (Tesar et al., 2003) test the ability of the algorithm to identify grammatical stress (most restrictive), lexical stress (least restrictive), and combinations of the two. The constraint set includes:

- MAINLEFT - Stress the leftmost syllable
- MAINRIGHT - Stress the rightmost syllable
- FAITHACCENT - Stress an accented syllable
- FAITHACCENTROOT - Stress an accented root syllable

Possible languages and their corresponding partial orders ranging from least restrictive to most restrictive are shown below. In the first two languages, the least restrictive languages, lexical distinctions in stress are realized faithfully, while grammatical stress surfaces only in forms with no underlying stress. In the final two languages stress is entirely grammatical; underlying distinctions are neutralized in favor of a regular surface stress pattern. Finally, the middle language is a combination of lexical and grammatical stress, requiring that the algorithm learn that a contrast in roots is preserved, while a contrast in suffixes is neutralized.

- Full contrast: roots and suffixes contrast in stress, default left:
  - F >> ML >> MR, FAR
  - pá$_1$-ka$_3$; pa$_1$-gá$_4$; bá$_2$-ka$_3$; bá$_2$-ga$_4$
- Full contrast: roots and suffixes contrast in stress, default right:
  - F >> MR >> ML, FAR
  - pa$_1$-ká$_3$; pa$_1$-gá$_4$; bá$_2$-ka$_3$; ba$_2$-gá$_4$
- Root contrast only, default right:
  - FAR >> MR >> ML
  - pa$_1$-ká$_3$; pa$_1$-gá$_4$; bá$_2$-ka$_3$; bá$_2$-ga$_4$
- Predictable left stress:
  - ML >> FAR, F, MR
  - pá$_1$-ka$_3$; pá$_1$-ga$_4$; bá$_2$-ka$_3$; bá$_2$-ga$_4$
- Predictable right stress:
  - MR >> FAR, F, ML
  - pa$_1$-ká$_3$; pa$_1$-gá$_4$; ba$_2$-ká$_3$; ba$_2$-gá$_4$

In all cases the algorithm learns the correct, restrictive grammars corresponding to the partial orders shown above. As before, the final lexicon assigns uniform probability to all underlying forms that differ in noncontrastive features. For example, in the case of the language with root contrast only, the final lexicon selects a unique lexical item for root morphemes and maintains a uniform probability distribution over stressed and unstressed underlying forms for suffixes.

4.3 Abstract Underlying Vowels

The final experiment tests the algorithm on an artificial language, based on Polish, with abstract underlying vowels that never surface faithfully. Although the particular phenomenon exhibited by Slavic alternating vowels is rare, the general phenomenon wherein underlying forms do not correspond to any surface allomorph is not uncommon and should be accommodated by the learning algorithm. This language presents a challenge for previous work on unsupervised learning of OT because alternations in the number of segments are observed in morpheme 3. The morphologically annotated input to the algorithm for this language is shown in Table 4.

kater$_1$	vatr$_2$	sater$_3$
kater$_1$-a$_4$	vatr$_2$-a$_4$	satr$_3$-a$_4$

Table 4. Yer Language Surface Forms

In this language morphemes 1, 2 and 4 exhibit no alternation while morpheme 3 alternates between *sater* and *satr* depending on the context. The constraints for this language, based on Jarosz (2005), are shown below:

- *E = *[+HIGH][-ATR]
- DEP-V
- MAX-V
- *COMPLEXCODA
- IDENT[HIGH]

1	2	3	4
/kater/	/vatr/	/satEr/	/-a/

Table 5. Desired Final Lexicon

In the proposed analysis of this language, the abstract underlying [E], which is a [+high] version of [e], is neutralized on the surface and exhibits two repairs systematically depending on the context. It deletes in general, but if a complex coda is at stake, the vowel surfaces as [e] by violating IDENT[HIGH]. The required partial ranking for this language is shown below while the desired lexicon is shown in Table 5.

{*E, {DEP-V >> *COMPLEXCODA }} >> IDENT[HIGH] >> MAX-V

The algorithm successfully learns the correct ranking above and the lexicon in Table 5. Specifically, the final grammar assigns equal probability to all the rankings consistent with the above partial order. The final lexicon selects a single underlying form for each morpheme as shown in Table 5 because all underlying distinctions in this language are contrastive.

4.4 Discussion

In summary, the algorithm is able to find a correct grammar and lexicon combination for all of the language systems discussed. As discussed in Section 3, the objective function itself does not favor restrictive grammars, but the ability of the algorithm to learn restrictive grammars in these experiments suggests that initializing the lexicons to uniform distributions does compel the learning algorithm to select restrictive grammars rather than restrictive lexicons.

While the experiments presented in this section focus on the task of learning a grammar and lexicon simultaneously, the proposed algorithm is also capable of learning grammars from structurally ambiguous forms. The same likelihood maximization procedure proposed here could be used for unsupervised learning of grammars that assign full structural description to overt forms. Future directions include testing the algorithm on language data of this sort.

5 Conclusion

In sum, this paper has presented an unsupervised, probabilistic algorithm for OT learning. The paper argues that combining the OT principle of Richness of the Base and likelihood maximization provides a novel and general solution to the problem of finding a restrictive grammar. The proposed solution involves explicitly implementing Richness of the Base in the initialization of the lexicon in order to fully utilize the properties of the objective function. By relying on Richness of the Base and likelihood maximization, the algorithm is able to use negative evidence implicitly to find restrictive grammars. The algorithm is shown to be successful on three constructed languages featuring different types of neutralization and hidden structure.

One potential extension of the proposed algorithm involves combining a system for unsupervised learning of morphological relations with the proposed algorithm for learning phonology. Several algorithms have been proposed for automatically inducing morphological relations, like those assumed by the present learner (Goldsmith, 2001; Snover and Brent, 2001). The task of uncovering morphological relations is complicated by allomorphic alternations that obscure the underlying identity of related morphemes. While these algorithms are very promising, their performance may be significantly enhanced if they were combined with an algorithm that models such phonological alternations.

In conclusion, this is the first proposed unsupervised algorithm for OT learning that takes advantage of the power of probabilistic modeling to learn a grammar and lexicon simultaneously. This paper demonstrates that combining OT theoretic principles with results from computational language learning is a worthwhile pursuit that may inform both disciplines. In this case the theoretical principle of Richness of the Base has provided a novel solution to a learning problem, but at the same time, this work also informs theoretical OT by providing a formal characterization of this theoretical principle. Future work includes testing on larger, more realistic languages, including language data with noise and variation, in order to determine the algorithm's resistance to noise and ability to model variable grammars like those observed in natural languages and in human language acquisition.

Acknowledgements

I would like to thank Paul Smolensky for his invaluable feedback on this work and for his suggestions on the preparation of this paper. I am also grateful to Luigi Burzio, Robert Frank, Jason Eisner, and members of the Johns Hopkins Linguistics Research Group (especially Joan Chen-Main, Adam Wayment, and Sara Finley) for additional comments and helpful discussion.

References

Apoussidou, Diana and Paul Boersma. 2004. Comparing Different Optimality-Theoretic Learning Algorithms:the Case of Metrical Phonology. *Proceedings of the 2004 Spring Symposium Series of the American Association for Artificial Intelligence.*

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R. Van Hout and W. L. Wetzels (eds.) Variation, Change and Phonological Theory. Amsterdam, John Benjamins.

Boersma, Paul. 1997. How we Learn Variation, Optionality, and Probability. *Proc. Institute of Phonetic Sciences of the University of Amsterdam* 21:43-58.

Boersma, P. 1998. *Functional Phonology*. Doctoral Dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, P. and B. Hayes. 2001. Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32(1):45-86.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from incomplete data via the EM Algorithm. *Journal of Royal Statistics Society*. 39(B):1-38

Eisner, Jason. 2000. Easy and hard constraint ranking in optimality theory: Algorithms and complexity. In Jason Eisner, Lauri Karttunen and Alain Thériault (eds.), *Finite-State Phonology: Proceedings of the 5th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 22-33, Luxembourg, August.

Escudero, Paola. 2005. *Linguistic Perception and Second Language Acquisition.Explaining the attainment of optimal phonological categorization*. Doctoral dissertation, Utrecht University.

Goldsmith, John. 2001. Unsupervised Learning of Morphology of a Natural Language. *Computational Linguistics*, 27: 153-198.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson and Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm University, pages 111-120.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. Appeared 2004 in Kager, Rene, Pater, Joe, and Zonneveld, Wim, (eds.), *Fixing Priorities: Constraints in Phonological Acquisition.* Cambridge University Press.

Jarosz, Gaja. 2005. Polish Yers and the Finer Structure of Output-Output Correspondence. *31st Annual Meeting of the Berkeley Linguistics Society,* Berkeley, California.

Lari, K. and S.J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*. 4:35-56

Legendre, Geraldine, Paul Hagstrom, Anne Vainikka and Marina Todorova. 2002. Partial Constraint Ordering in Child French Syntax. to appear *Language Acquisition* 10(3). 189-227.

Merchant, Nazarré, and Bruce Tesar. to appear. Learning underlying forms by searching restricted lexical subspaces. In *The Proceedings of Chicago Linguistics Society* 41. ROA-811.

Pereira, F. and Y. Schabes. 1992. Inside-Outside reestimation from partially bracketed corpora. In *Proceedings of the ACL 1992*, Newark, Delaware.

Prince, Alan and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report 2, Center for Cognitive Science, Rutgers University.

Prince, Alan, and Bruce Tesar. 1999. Learning phonotactic distributions. Technical Report RuCCS-TR-54, Rutgers Center for Cognitive Science, Rutgers University.

Riggle, Jason. 2004. Generation, Recognition, and Learning in Finite State Optimality Theory. Ph.D. Dissertation, UCLA, Los Angeles, California.

Rosenbach, Anette and Gerhard Jaeger. 2003. Cumulativity in Variation: testing different versions of Stochastic OT empirically. Presented at the *Seventh Workshop on Optimality Theoretic Syntax*, University of Nijmegen.

Smolensky, Paul. 1996. The initial state and `richness of the base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Department of Cognitive Science, Johns Hopkins University.

Snover, Matthew and Michael R. Brent. 2001 A Bayesian Model for Morpheme and Paradigm Identification. In *Proceedings of the 39th Annual Meeting of the ACL*, pages 482-490. Association for Computational Linguistics.

Tesar, Bruce. 1995. *Computational Optimality Theory*. Ph.D. thesis, University of Colorado at Boulder, June.

Tesar, Bruce. 1998. An iterative strategy for language learning. *Lingua* 104:131-145. ROA-177.

Tesar, Bruce. 1999. Robust interpretive parsing in metrical stress theory. In *The Proceedings of Seventeenth West Coast Conference on Formal Linguistics*, pp. 625-639. ROA-262.

Tesar, Bruce. 2004. Contrast analysis in phonological learning. Manuscript, Linguistics Dept., Rutgers University. ROA-695.

Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani, and Alan Prince. 2003. “Surgery in language learning”. In *The Proceedings of Twenty-Second West Coast Conference on Formal Linguistics*, pp. 477-490. ROA-619.

Tesar, Bruce and Alan Prince. to appear. “Using phonotactics to learn phonological alternations.” Revised version will appear in *The Proceedings of CLS 39*, Vol. II: The Panels. ROA-620.

Tesar, Bruce and Paul Smolensky. 1995. “The Learnability of Optimality Theory”. In *Proceedings of the Thirteenth West Coast Conference on Formal Linguistics*, 122-137.