# Using Entropy to Learn OT Grammars From Surface Forms Alone

**Jason Riggle**
**University of Chicago**

## 1. Introduction

The problem of ranking a set of constraints in Optimality Theory (Prince and Smolensky 1993) in a fashion that is consistent with an observed training sample comprised of ⟨input, output⟩ pairs has been solved with a variety of algorithms (e.g. Tesar 1995; Tesar and Smolensky 1993, 1996, 1998, 2002; Boersma 1997; Boersma and Hayes 2001). In real-world language learning scenarios, however, learners aren't usually presented with ⟨input, output⟩ pairs but must instead learn grammars from output forms alone (possibly aided by conjectures about their meanings or morphology).

The task of learning OT grammars from training samples that consist of output forms alone presents many challenges. Chief among these is the problem that there are often many ⟨possible-input, possible-grammar⟩ pairs that are consistent with any observed training sample of surface forms. For instance, the fully faithful 'identity grammar' is a perennially viable hypothesis under which the input forms are assumed to be basically identical to the observed output forms. In addition to this hypothesis, depending on the particular constraints in CON and the specific surface forms in the training sample, there are a potentially huge number of other ⟨grammar, i/o-mapping-set⟩ pairs, each one deriving a different set of unfaithful mappings from potential input forms to the outputs in the training sample.

Knowledge of meaning and morphology can help the learner choose grammar hypotheses that map the same input to different surface instantiations of the same morpheme. But, even before the learner has any knowledge about the morphology and meanings of words, it is possible to make educated guesses about the structure of the grammar. Suggested strategies for this include principles like Prince and Tesar's (1999) selectional preference for ranking hypotheses that are maximally 'restrictive' or Smolensky's (1996) default MARKEDNESS >> FAITHFULNESS ranking. Both of these strategies restrict the search through the space of possible grammars by providing a heuristic that's designed to prefer rankings that generate a tight fit with the observed data.

In this paper I present yet another strategy for adjudicating among competing grammar hypotheses without recourse to morphological information or meanings. The strategy that I propose is not based on formal properties of the constraint rankings themselves, but instead is based on information-theoretic properties of the set of inputs that each candidate grammar (ranking) associates with the training sample. The idea is this: if learners choose grammars whose associated input sets have the highest entropy (are least ordered) they will select grammars that maximally characterize patterns in the training sample as consequences of the grammar rather than as accidents of the lexicon.

## 2. Entropy, restrictiveness, and Richness of the Base

The idea that gaps in the attested phonological patterns of a language, especially those involving elements that are typologically marked, should be encoded in the grammatical description of the language is far from new. It is also clear that speakers are sensitive not only to total gaps but also (at some level) to rather subtle stochastic generalizations about what sorts of structures are 'good' (or perhaps simply typical) in their language (cf. Pierrehumbert, Beckman and Ladd 2001, and references

© 2006 Jason Riggle. Proceedings of the 25th West Coast Conference on Formal Linguistics, ed. Donald Baumer, David Montero, and Michael Scanlon, 346-353. Somerville, MA: Cascadilla Proceedings Project.

therein). In derivational models like Chomsky and Halle's (1968), these sorts of patterns can either be directly encoded via Morpheme Structure Constraints or they can arise via rule interaction. In Optimality Theory, the Richness of the Base Hypothesis (ROTB; Smolensky 1996) demands that these sorts of patterns be generated by constraint interaction. Though these models differ in substantive ways, the idea that the phonological grammar should encode patterns (even if they are also present in the lexicon in rule based models) is common to both.

The idea that the types of patterns the grammar must account for are mathematical deviations from simple uniform (or random) distributions is central in Zellig Harris' work (1942 *et seq.*). What I propose here is a method for adjudicating among competing grammar hypotheses that uses the idea that grammatical patterns are information set against a background of randomness to select hypotheses that maximally encode observed patterns as consequences of the grammar. To do this we derive for each grammar hypothesis the set of input forms that it maps to the observed output forms and then evaluate each input set to find the grammar that maps the least restricted (most random) set of inputs to the observations. When successful, this method will ensure that only idiosyncratic information is part of the input set while all patterns that can be derived by the grammar are part of the grammar.

This strategy is very general and could be used for any phonological model in which it is possible to enumerate a range of grammars consistent with a training sample and derive for each one the set of inputs that it associates with the training sample. The actual implementation of this strategy, on the other hand, can be quite complicated and requires myriad representational and computational decisions that are specific to the phonological model in which it is to be used. Henceforth I will restrict my attention to Optimality Theory where the hypothesis of a fixed universal constraint set makes it (relatively) easy to enumerate and evaluate the competing grammar hypotheses.

In OT, the heuristic of selecting grammars with the most entropic input sets can be seen as a direct implementation of what Smolensky (1996:1) describes as "... a fundamental OT principle, *Richness of the Base*: the set of possible inputs to the grammar is universal." This is sometimes paraphrased as an edict forbidding constraints on inputs and sometimes maligned as a principle that does not allow for a substantive theory of the lexicon or a nuanced theory of language variation. For the most part, these criticisms misapprehend the architecture of OT. Given a universal constraint set CON, ROTB simply says that if faced with a decision whether to encode a phonological pattern as a consequence of constraint interaction or as an accident of the lexicon the former option must be taken. This still leaves all idiosyncratic information (which can include patterns that aren't derived by the constraints in CON) in the lexicon. ROTB is not antithetical to models of speaker's knowledge of their lexicons, it simply says that if a pattern can be in the grammar it should be.

## 3. Entropy as a window on the input/output relation

The idea that learners should encode gaps in the phonological patterns they observe as part of their grammar is an old one. Learning the phonotactics of a language can proceed before the learner knows meanings or morphology because the learner needn't know what the inputs are to notice the gaps in the observed output forms. In Optimality Theory, however, it is possible to exploit the assumption that all learners have access to the universal constraint set CON to go beyond phonotactic learning and make educated guesses about the input/output relation even when the learner only has access to the outputs.

Because the constraints are universal, the learner knows (a) what sorts of marked structures might be eliminated under various rankings and (b) what sorts of unmarked structures might arise through different repair strategies under other rankings. If the learner can safely make some assumptions about the distribution of input structures, e.g. that they are basically equiprobable, or that they obey Zipf's (1932) Law, or that they decrease in probability with increasing complexity, etc., then various repairs of marked structures will leave telltale perturbations in the distribution of structures among surface forms. I will discuss the specific probabilistic models in section 4. For now, let's make the assumption that input structures are basically equiprobable and consider an example of how four languages might deal with coda consonants.

| | | |
|---|---|---|
| (1) | $L^1$: tolerates codas | Result: surface distribution unaffected |
| | $L^2$: deletes heterorganic codas | Result: all codas are homorganic |
| | $L^3$: deletes all codas | Result: total absence of codas |
| | $L^4$: epenthesizes [i] creating a new syllable | Result: abundance of [Ci] syllables, no codas |

If the learner has four grammars to chose from where $G^1$ is the identity grammar, $G^2$ deletes codas that are not homorganic, $G^3$ deletes codas, and $G^4$ epenthesizes [i] to turn codas into onsets, then how will each grammar characterize samples from the various languages? In the simplest case, a sample from $L^1$ that contains a heterorganic coda will be incompatible with $G^2$, $G^2$, and $G^4$ thereby allowing only one grammar to be associated with the sample. Samples from $L^{2\text{-}4}$ are more interesting. A sample from $L^2$ could be paired with either $G^1$ or $G^2$ while a pairing $G^3$ or $G^4$ is impossible due to the presence of codas. The input set assigned by $G^1$ will have a lot of structure – all codas will just happen to be homorganic, and non-homorganic codas will just happen to be absent – while the input set assigned by $G^2$ will not contain this kind of structure. In this case selecting the grammar with the more entropic (i.e. less structured) input set correctly identifies the language. A sample of $L^3$ can be paired with any of the grammars. Nonetheless, the input set for $G^1$ will contain structure – the absence of codas, the input set for $G^2$ will contain structure – the absence of homorganic codas, and the input set for $G^4$ will contain structure – every coda-less syllable followed by a syllable without an [i] nucleus will have to be coda-less in the input. The correct grammar $G^3$, on the other hand, will not put this kind of structure in the input set and will thus have the most entropic input set.

From this description it is obvious that an accidental perturbation in the distribution of surface forms can easily trick the learner into choosing the wrong grammar. This is an asset of the model. This will make it possible to tune the learner's sensitivity subtle to probability perturbations so as to match the type of judgments that humans give when presented with training data that contains statistical patterns. The best fit with human performance in this regard could be achieved if, instead of categorically choosing among grammars, the learner assigned probabilities to each grammar. This scenario might be implemented with Boersma and Hayes' (2001) stochastic model of OT, but in order to do this it would be necessary to render the range of grammars discrete, so that input sets could be calculated for each possible grammar. This is an interesting problem but beyond the scope of this paper. For the remainder of paper I will assume that the OT grammars the learner uses are non-stochastic.

Stepping back from the specifics of the computation of the entropy of input sets, it is accurate to describe the goal of the entropy-based learner as mapping the observed output forms to the broadest possible cover over the infinite range of potential input forms. This strategy is only a heuristic, it can be led astray by accidental patterns, and yet still other properties of the input/output relation might not yield a unique or discernable perturbation in the surface forms. This strategy should be seen as a way of bootstrapping the learner into some grammar hypotheses based on surface patterns that would be used in tandem with other converging lines of evidence from meaning and morphology once they are available.

## 4. Using entropy in practice

The biggest hurdles in implementing the entropy-based leaner are (i) finding a safe assumption about the baseline probability of input structures, (ii) sorting, updating, and storing a potentially huge number of ⟨grammar, input-set⟩ pairs in a realistic on-line fashion that doesn't require unreasonable amounts of memory, and (iii) inverting the phonological grammar to generate ⟨grammar, input⟩ pairs from surface forms.

The problem of finding reasonable assumptions about the baseline probability of input structures is an empirical one. The easiest things for the learner to observe are the frequencies of structures in the running text of observed surface forms. However, given Zipf's (1934) observation that the frequency of a word (or morpheme, or n-gram) in a text is approximately inversely proportional to its rank in a frequency table, there is the danger that high-frequency surface forms are prevalent for reasons that might have little to do with the phonological grammar and thus are not relevant in assessing what sorts of repair strategies the language employs to fix marked structures. Using a set of surface form tokens is a better strategy but has its own problems in that it requires the learner to remember a set of surface forms and, in languages with complex morphology, a set of observed 'words' may contain repetitions of various morphemes and thus introduce some of the same distributional properties of texts.

For the test case discussed in section 7, I counted the frequency of structures in the legal tokens of the various languages following the assumptions of Frisch, Broe and Pierrehumbert (1997) that, all else equal, the various structures should be equally represented in legal words of a language. The assumption that input structures should be roughly equiprobable is not crucial. It is only necessary that the learner have some (reasonably correct) characterization of the likely distribution of input forms. Even if this is not a reasonable assumption about human language, the equiprobability hypothesis (or some variant) might still be an effective heuristic in the early stages of phonological acquisition to give the learner a place to start in searching through the range of grammar hypotheses. Clearly more real acquisition data must be brought to bear on this issue.

## 5. Managing the hypotheses

This is one of the biggest problems in scaling up from toy grammars to real grammars. Even with the ten constraints discussed in section 7 there are about 3.6 million rankings and as such the number of ⟨grammar, i/o-mapping-set⟩ pairs consistent with each sample of surface forms can quickly climb into the hundreds. In considering the added complication that each i/o-mapping-set can be as large and complex as the sample of surface forms it quickly becomes obvious that it is not feasible to work with full descriptions of every input set that can be mapped to the observed surface forms. To overcome these problems I implemented two solutions. First, rather than keeping an entire input set with each grammar hypothesis I kept only a table of counts for pairs of adjacent segments. This data can then at any point be cashed out as a table of bigram frequencies which can be used to assess the entropy of the input set. Second, rather than keeping all the grammar hypotheses, I kept only the ***n*** hypotheses whose input sets had the highest entropy (where ***n*** was usually set to 10).

These simplifications run the risk of introducing two types of errors. First, the reduction of the input sets to bigrams will render invisible any structures that are articulated over more than two adjacent segments (e.g. vowel harmony). While this was not a concern for the constraint set in my test case it is probably not a good general solution. There are encoding schemes that are capable of the same sort of reduction of the input set I obtained with bigrams (e.g. Markov random fields) that can encode dependencies among non-adjacent elements. However, any such encoding scheme other than recording the entire input set will lose some information. Crucially, we do not need a perfect encoding scheme, but merely one that has the same sort of sensitivity that human learners have. In spite of its limitations, a bigram model seems like a good start in this regard.

The second simplification, keeping only the ***n*** hypotheses with the highest entropy, has interesting consequences. In monitoring the performance of my learner through several test cases it became clear that the 'correct' hypothesis could readily be knocked out of the Top 10 by other incorrect hypotheses with higher entropy. Because I presented randomly generated training data to the learner, this occurred whenever the randomly drawn data was consistent with a surface language that allowed a subset of the forms of the actual target language. In these cases there were often myriad grammar hypotheses that mapped the training sample to a larger and less ordered input set than the target grammar. Though, it might initially seem like a flaw, this type of overgeneralization is very human-like. Moreover, this is precisely the type of error that can be recovered from. When an entropy-based learner is presented with training data that is consistent with a language that is a subset of the target language the learner will always pick the subset language, but all it takes is a future observation of a datum outside the subset to move the learner to the larger language. In this way the entropy heuristic is an implementation of the all-important subset principle in learning (Baker 1979, Pinker 1979, Dell 1981, Manzini and Wexler 1987, Clark 1992, and many others).

## 6. Inverting the grammar

The task of inverting OT grammars presents many technical challenges. This topic is worthy of lengthy discussion by itself but here I will only be able briefly touch on some of the issues that arise and a couple of strategies for grammar inversion. The problem of grammar inversion is separate from any strategies for using the information gleaned from the inverted grammars but must be solved (at least approximately) for any learning from surface forms to be possible. The learning algorithm that I present in this work simply assumes that grammar inversion is possible and that learners have access to the inputs that are mapped to observed surface forms by the range of possible grammars.

The best way to truly invert OT grammars would be to try to emulate Kaplan and Kay's (1994) computational characterization of rule-based phonological grammars as a set of composed transducers each one implementing a single ordered rule. Unfortunately this is not easy in OT and in some cases it is impossible. Frank and Satta (1998) showed that, in general, there are optimizations in OT that cannot be characterized with finite-state transducers. There have been a variety of proposals for getting around this obstacle – e.g. Karttunen (1998) proposes bounding the number of constraint violations, Eisner (2000) proposes directional evaluation of constraints, Gerdemann and Van Noord (2000) provide a way permute and match up violations in competing candidates, and Riggle (2004) provides a transducer construction scheme that crashes for grammars that are not finite-state. All these strategies, however, only generate a single input-to-output transducer for a specific ranking. What is needed is a general transducer that can simultaneously generate optimal input/output pairs for all rankings.

Initial investigations for a few small constraint sets (including the one in this paper) suggest that it is possible to combine the algorithm for generating contenders (candidates that are not harmonically bounded) given in Riggle (2004) with the algorithm for turning sets of OT constraints into transducers. The process is computationally expensive – for the 10 constraints in section 7 the compilation time for the transducer was several days – so it may not be a generally feasible strategy, but it is worth further investigation.

An alternative solution that is much easier to implement is to simply use a finite range of input forms in the investigation. By using the CONTENDERS algorithm from Riggle (2004) it is possible to generate for each input form a range of candidates that are not harmonically bounded. Comparing each contender with its competitors yields for each one a set of partial rankings under which it is optimal. The partial rankings arguments can be succinctly encoded with Prince's (2002) Elementary Ranking Conditions (ERCs). With a finite input set it is then possible to generate by brute force a table of ⟨input, output, ERCs⟩ triples. Though it is inelegant, this table can be used as an oracle to supply the learner with the ⟨grammar, input⟩ pairs associated with an observed surface form.

## 7. An illustration with syllable theory

Prince & Smolensky's (1993) syllable structure grammar with an alphabet of two consonants and two vowels can provide a small test-case for the strategy. If, under some rankings, one of the vowels or one of the consonants is relatively less marked than the other then this asymmetry can be reflected in the repair strategies that are used in the various possible languages.

(2) Phoneme inventory: {b, p, i, a}

(3) Ten constraints

MAX V: penalizes deletion of vowels
MAX C: penalizes deletion of consonants
DEP V: penalizes insertion of vowels
DEP C: penalizes insertion of consonants
IDENT(HEIGHT): penalizes changes in height
IDENT(VOICE): penalizes changes in voicing

ONSET: penalizes syllables without onsets
NOCODA: penalizes syllables with codas
*VOICED OBST: penalizes voiced obstruents
*HIGH V: penalizes high vowels

For the purposes of this examination, I assume that the surface phoneme inventory is the same as the underlying inventory, that all input strings are unsyllabified, that undominated markedness constraints demand that all output segments be syllabified in syllables with exactly one nuclear vowel and no consonant clusters, and that undominated faithfulness constraints forbid fission, fusion, metathesis, and changes of vowels into consonants. With this inventory, this constraint set, and this set of assumptions we are ready to simulate learning.

For the set of inputs in the test case I used the 340 strings between one and four segments long that can be generated from the alphabet {b, p, i, a}. For each of the input strings I generated the range of candidates that contenders and from each ⟨input, contender⟩ pair I generated an ⟨input, contender, ERCs⟩ triple where ERCs refers to the Elementary Ranking Conditions (Prince 2002) that define a disjunction of partial constraint rankings under which that contender is more harmonic than the other contenders for the same input. An example contender set is given in (4).

(4) Nineteen contenders for input /abba/:

| /abba/ | DEP C | DEP V | ID(HT) | ID(VOI) | MAX C | MAX V | *HIGH V | *VOI OBST | NOCODA | ONSET |
|---|---|---|---|---|---|---|---|---|---|---|
| a. .ab.ba. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 |
| b. .a.ba. | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| c. .ba. | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| d. .a.a. | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| e. ∅ | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| f. .a.pa. | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| g. .pa. | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| h. .ap.pa. | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 1 |
| j. .a.ba.ba. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 |
| k. .ba.ba. | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| l. .a.pa.pa. | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |
| m. .pa.pa. | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| n. .pab.ba. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| o. .pa.ba. | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| p. .pa.pa. | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| q. .pap.pa. | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| r. .pa.ba.ba. | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| s. .pa.pa.pa. | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| t. .pa.pa. | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |

The input /abba/ yields 19 ⟨input, output, ERC⟩ triples. In total, the 340 input strings in the test case produced 5,417 ⟨input, output, ERC⟩ triples. Among this data there were 573 unique output forms. It is possible to compute the typology realized by this constraint set over these inputs by doing a depth-first search in which one ⟨output, ERC⟩ for each of the 340 inputs are combined provided that the union of the ERCs in the pairs are not internally contradictory. Carrying this out for the 340 input strings reveals 330 unique languages (where a language is a set of input/output mappings) which correspond to 120 unique surface-languages (where the surface-language) is just the output forms. This is a far cry from the upper bound of 3,628,800 languages that could be generated under factorial reranking of the constraints.

For the test I randomly selected one of the 340 languages and then began presenting the learning algorithm with surface forms drawn randomly from that language's set of licit surface forms under a uniform distribution. The entropy learner can be characterized in very loose pseudo-code as in (5).

(5) The entropy learner

```
1   The learner starts with an empty set of hypotheses $\boldsymbol{H} = \{(\varnothing,\varnothing)\}$.
2   The learner starts with a empty set of observed forms $\boldsymbol{Obs} = \varnothing$.
3   The teacher presents a surface form $\alpha$ to the learner.
4       If the $\alpha$ is in Obs the learner returns to 3 for another example
5       else the learner obtains from the Oracle the set $\Phi$ of $\langle$output, ERC$\rangle$ pairs for $\alpha$
6           for each $(o, e)$ pair in $\Phi$ and
7               for each hypothesis $h = (o', e')$ in $\boldsymbol{H}$
8                   if $e \cup e'$ is an internally consistent set of ranking statements
9                       then add $(o \cup o', e \cup e')$ to $\boldsymbol{H}$
10      the learner then discards all but the 10 hypotheses in $H$ whose input sets have the highest
        entropy as computed in (6) below and returns to 3 for another example
```

To evaluate the entropy of the input-sets in the competing hypotheses. I represented each one as a set of bigram counts which were then, at the point of evaluation, turned into frequencies. For this study I used conditional entropy to get at patterns and mutual information among pairs of adjacent segments. The formula for computing conditional entropy is given in (6).

(6) $$H(Y \mid X) = -\sum_{x \in \Sigma} \sum_{y \in \Sigma} p(x, y) \log p(y \mid x)$$

In this formula we take the inverse of the sum of probability of each pair of segments in the data set $p(x, y)$ multiplied by the base-2 log of the conditional probability of $y$ given $x$.

Using this algorithm I ran 100 trials assessing for each one the number of randomly drawn samples required to get the learner to a grammar that generated the set of input/output mappings in the teacher's language. The average number of samples required to converge on a correct surface language was 258 with a wide range of variation between the trials. I should note that I halted two trials where the learner seemed to get perpetually stuck. Whether these trials merely represent the tail of the probability curve for the number of samples required, were the result of a bug in the programming, or represent a potential trap in the traversal of the hypothesis space is still under investigation.

## 8. Conclusions and extensions

Though these results are for an extremely simple 'toy' grammar they are quite promising. It seems that there may be a wealth of information hiding in the surface distribution of forms that the learner could exploit to get a handle on the grammar.

The single biggest outstanding issue for this proposal is whether the assumption that various input structures are equiprobable (or have some knowable distribution) is workable in the real world. This will only be resolved with further empirical tests of the learning strategy. It is worth noting here that Jarosz (2006) has gotten good results in unsupervised OT learning with the same sorts of assumptions, but in a mirror image of the proposal here – put loosely, Jarosz starts by assuming that the input-set is maximally entropic and then finds the grammar that best fits this scenario.

The biggest outstanding technical hurdle is the need for an efficient strategy for inverting OT grammars. Though the look-up table implemented for the Oracle was adequate for this small study, this method will not work for large (or infinite) ranges of inputs and for large grammars.

Assuming that these issues can be overcome, and that a good succinct representation of the input sets (either in terms of bigrams, or some other simplified model) is available, this strategy offers a way to implement the subset principle that does not require learners to have an innate knowledge of when one grammar generates a subset of the forms generated by another grammar.

Though the Richness of the Base assumption was crucially exploited in the process of choosing among competing grammar hypotheses to link disparities in the prevalence of surface structures to the grammar (rather than the inputs), it need not be the case that the actual lexicon that the learner ends up with be free of structure. It is entirely feasible that this strategy could grammaticize patterns in the training data that ultimately get duplicated in the adult lexicon.

## References

Baker, Charles. L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10, 233-280.S

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

Boersma, Paul and Bruce Hayes 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45-86.

Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* 2.2, 83-149.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12.1, 31-37.

Eisner, Jason. 2000. Directional constraint evaluation in Optimality Theory. In *Proceedings of COLING*.

Frank, Robert and Giorgio Satta. 1998. *Optimality Theory and the Generative Complexity of Constraint Violability*, Computational Linguistics 24, 307--315.

Frisch, Stefan A., Michael B. Broe and Janet B. Pierrehumbert. 1997. ‘Similarity and Phonotactics in Arabic’, unpublished manuscript, Indiana University

Gerdemann, Dale, and Gertjan v. Noord. 2000. Approximation and exactness in finite state optimality theory. In Jason Eisner, Lauri Karttunen, and Alain Thriault (eds.), *SIGPHON 2000, Finite State Phonology*.

Jarosz, Gaja. 2006. A Probabilistic Unsupervised Algorithm for Learning Optimality Theoretic Grammars. Presentation at the *80th Annual Meeting of the Linguistic Society of America*, Albuquerque, New Mexico.

Kaplan, Ronald M. and Martin Kay. 1994. Regular model of phonological rule systems. Computational Linguistics, 20(3):331-378.

Karttunen, Lauri. 1998. The proper treatment of optimality theory in computational phonology. In *Finite-state Methods in Natural Language Processing.* 1–12.

Manzini, M. Rita and Wexler, Kenneth. 1987. Parameters, binding theory, and learnability. *Linguistic Inquiry* 18.3, 413-444.

Merchant, Nazarré, and Bruce Tesar. 2006. Learning underlying forms by searching restricted lexical subspaces. In *The Proceedings of CLS 41*.

Pierrehumbert, Janet, Mary Beckman, and D. Ladd. 2001. Conceptual Foundations of Phonology as a Laboratory Science. In Burton-Roberts, N., Carr, P. and Docherty, G. (eds) Phonological Knowledge, Oxford, UK: Oxford University Press, 273-304.

Pinker, Steven. 1979. Formal models of language learning. *Cognition*, *7*, 217-283.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: constraint interaction in generative grammar. TR-2, Rutgers University Cognitive Science Center.

Prince, Alan S. 2002. Entailed Ranking Arguments. ROA-500.

Riggle, Jason 2004. *Generation Recognition and Learning in Finite State Optimality Theory*. PhD Dissertation, University of California, Los Angeles.

Tesar, Bruce. 1995. *Computational Optimality Theory*. Doctoral dissertation, University of Colorado.

Tesar, Bruce, 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, University of Colorado at Boulder. Rutgers Optimality Archive ROA-2, http://ruccs.rutgers.edu/roa.html.

Tesar, Bruce, and Paul Smolensky. 1996. *Learnability in Optimality Theory (long version).* Technical Report 96-3, Department of Cognitive Science, Johns Hopkins University, Baltimore. ROA-156

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press

Zipf, George. Kingsley. 1932. Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA: Harvard University Press.

# Proceedings of the 25th West Coast Conference on Formal Linguistics

edited by Donald Baumer,
David Montero, and Michael Scanlon

Cascadilla Proceedings Project     Somerville, MA     2006

## Copyright information

ISBN 1-57473-415-6 library binding

## Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

## Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Riggle, Jason. 2006. Using Entropy to Learn OT Grammars from Surface Forms Alone. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. Donald Baumer, David Montero, and Michael Scanlon, 346-353. Somerville, MA: Cascadilla Proceedings Project.

or:

Riggle, Jason. 2006. Using Entropy to Learn OT Grammars from Surface Forms Alone. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. Donald Baumer, David Montero, and Michael Scanlon, 346-353. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1467.