A Maximum Entropy Model of Phonotactics and Phonotactic Learning*

Bruce Hayes Colin Wilson

UCLA

July 2006

1. Introduction

In one of the central articles from the early history of generative phonology, Chomsky and Halle (1965) lay out a research program for the theory of phonotactics. They begin with the observation that the logically possible sequences of English phonemes can be divided into three categories:

(1) a. Existing words, such as *brick*
b. Non-existing words that are judged by native speakers to be well-formed, such as *blick*.
c. Non-existing words that are judged by native speakers to be ill-formed, such as *bnick*.

The scientific challenge posed by this categorization has two parts. The first is to characterize the grammatical knowledge that permits native speakers to make phonotactic well-formedness judgments. The second, more fundamental challenge is to understand the principles with which phonotactic grammars are acquired.

The difficulty of this task is evident from a second point made by Chomsky and Halle, namely that there are grammars that fully cover the learning data but fail to capture the native speaker's knowledge. For example, they note (p. 101) that both of the rules given in (2) are compatible with the data available to the learner. However, while the rule in (2a) correctly excludes **bnick*, it also excludes the acceptable form *blick*. In contrast, (2b) appropriately excludes **bnick* but allows *blick* as a possible word.

(2) a. Consonantal Segment → r / # b___ɪk
b. Consonantal Segment → Liquid / # Stop ___ Vowel

The problem of phonotactic learning, then, is that of *selecting* a particular grammar — the one that is in fact acquired by native speakers — from among all of the possible grammars that are compatible with the learning data. Chomsky and Halle schematize the selection process as follows, where "AM" is the universal mechanism, or "acquisition model," that projects grammars from data.

* We would like to thank Steven Abney, Jason Eisner, Robert Malouf, Donca Steriade, and audiences at the University of Michigan, University of California at San Diego, and UCLA for helpful input on our project.

(3) primary linguistic data → AM → grammar

In this paper, we take up the challenge posed by Chomsky and Halle, proposing an explicit theory of phonotactic grammars and of how those grammars are learned. We propose that phonotactic grammars are composed of numerically-weighted constraints, and that the well-formedness of an output is determined by the weighted sum of its constraint violations. We further propose a learning model in which constraints are selected from a constraint space provided by UG and assigned weights according to the principle of *maximum entropy*. This model learns phonotactic grammars from representative sets of surface forms. In this paper, we apply the model to data from a number of languages, including English, Shona, and Wargamay. We show that the learned grammars capture the distributional generalizations of the languages and accurately predict experimental findings.

The organization of the paper is as follows. §2 elaborates our research goals in constructing a phonotactic learner, while §3 and §4 describe our learning model in detail. The next four sections are case studies, covering English syllable onsets (§5), Shona vowel harmony (§6), stress systems (§7), and finally a whole-language analysis, Wargamay (§8). In the concluding section (§9), we address questions raised by our work and outline directions for future research.

2. Goals of a phonotactic learner

We claim that the following criteria are appropriate for evaluating theories of phonotactics and phonotactic learning.

2.1 Expressiveness

The findings of the last few decades demonstrate a striking richness of structures and phenomena in phonology, including long-distance dependencies (e.g., McCarthy 1988), phrasal hierarchies (Selkirk 1980a), metrical hierarchies (Liberman and Prince 1977, Prince 1983), elaborate interactions with morphology (Kiparsky 1982), and other areas, each the subject of extensive analysis and research. We anticipate that a successful model of phonotactics and phonotactic learning will incorporate theoretical work from all of these areas.

If this view is correct, it follows that any approach to phonotactics in which the content of constraints is hobbled by computational considerations should be rejected. Thus, for instance, traditional *n*-gram models (Jelinek 1999, Jurafsky and Martin 2000) are quite efficient and have broad application in industry, but are insufficient as a basis for phonotactic analysis (see §5.3). Similarly, the stochastic context-free grammar of Coleman and Pierrehumbert (1997), while more phonologically sophisticated, rests on a single partition of words into onsets and rimes. As Coleman and Pierrehumbert point out, this makes it impossible in principle for the model to capture the many phonotactic restrictions that cross onset-rime boundaries (Clements and Keyser 1983:20- 21) or syllable boundaries (bans on geminates, heterorganic nasal-stop clusters, sibilant clusters). The crucial point is that phonotactics are cross-classifying, so that no one single categorization can encompass them all. Thus we believe it is necessary to adopt a model that fully engages with phonological theory. The theory of maximum entropy provides the desired level of expressiveness.

2.2 Providing an inductive baseline

While we have emphasized the primacy of phonological theory, the precise content of the latter remains an area of considerable disagreement. A computational learning model can be used as a tool for evaluating and testing theoretical proposals. The idea is that a very simple theory can provide a sort of *inductive baseline* against which more advanced theories can be compared. If the introduction of a theoretical concept makes possible the learning of phonotactic patterns that are inaccessible to the primitive baseline system, the concept is thereby supported. For earlier work pursuing the inductive baseline approach, see Gildea and Jurafsky 1996, Peperkamp et al. (in press).

Our own inductive baseline is a purely linear, feature-bundle approach modeled on Chomsky and Halle (1968; henceforth *SPE*). To this we will add the concepts of autosegmental tier (Goldsmith 1979) and metrical grid (Liberman 1975, Prince 1983), showing that both make possible modes of phonotactic learning that are unreachable by the linear baseline model.

2.3 Accounting for gradience

All areas of generative grammar that address well-formedness are faced with the problem of accounting for gradient intuitions. A large body of research in generative linguistics deals with this issue; for example Chomsky 1963; Ross 1972; Legendre et al. 1990; Schütze 1996; Hayes 2000; Boersma and Hayes 2001; Boersma 2004; Keller 2005; Sorace and Keller 2005; Legendre et al. 2006. In the particular domain of phonotactics, gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale (e.g., Greenberg and Jenkins 1964, Ohala and Ohala 1986, Coleman and Pierrehumbert 1997, Vitevitch et al. 1997, Frisch et al. 2000, Treiman et al. 2000; Bailey and Hahn 2001, Hay, Pierrehumbert, and Beckman 2003, Hammond 2004) and binary responses yield similar responses when averaged across participants (Coleman and Pierrehumbert 1997, Pierrehumbert 1994, Pater and Coetzee 2006). Thus, we consider the ability to model gradient intuitions to be an important criterion for evaluating phonotactic models. As we will show below, it is an inherent property of maximum-entropy models that they can account for both categorical and gradient phonotactics in a natural way.

To sum up, we seek to solve Chomsky and Halle's problem, specifying the structure of the module AM and testing it on actual phonotactic systems, with the goal of describing the full range of data including gradient intuitions. As a research strategy, we adopt the inductive baseline approach, requiring that phonological theories justify themselves through improvements in learning performance. To this end, we adopt an overall framework for learning, maximum entropy, that is neutral with regard to the constraints employed. We turn next to the structure of this model.

3. Maximum entropy

In this section, we justify maximum entropy on general grounds as the basis for a phonotactic learning model, then cover the basic concepts. For a general introduction to this theory, see Jelinek (1999:ch. 13), and for application to the learning and analysis of input-output mappings, see Goldwater and Johnson 2003, Jäger 2004.

Maximum entropy grammars ("maxent"; Berger et al. 1996, Della Pietra et al. 1997, Eisner 2001, Klein and Manning 2003, Rosenfeld 1996) have special properties that recommend them as a basis for phonotactic learning. In particular, they have been subject to thorough mathematical analysis, analysis that establishes their convergence properties and connection to the theories of information and statistical estimation. Moreover, the solutions they embody can be said to have a highly principled character, discussed in more detail below.

3.1 Constraint weighting

A maxent grammar consists of a set of constraints that are stated in the chosen representational vocabulary.[1] The constraints are free to refer to all of the featural, structural, and other distinctions made by the representations, and thus are not restricted by any specific predetermined categorization such as into syllabic roles. Any constraint that could be written in Optimality Theory ("OT"; Prince and Smolensky 1993/2004) or other constraint-based grammatical formalisms could also be a maxent constraint.

Maxent models are distinguished not by the content of their constraints but by the method of constraint interaction that they employ. The constraints are not taken to be inviolable (and they may in fact be violated by many legal forms), nor is their interaction regulated by a strict-domination hierarchy (as in OT). Instead, each constraint has an associated strength, or *weight*. The weight is a real number that can be thought of as scaling the importance of one constraint relative to others in the same grammar. The constraints and their weights determine the relative well-formedness of a representation as stated in definition (4).

(4) *Definition: Maxent value*

Given:

- a grammar with constraints C_1, C_1, …, C_N,
- associated weights w_1, w_2, … w_N, and
- a phonological representation x,

the *maxent value* of x, denoted $h(x)$, is:

$$\exp(-\sum_{i=1}^{N} w_i C_i(x))$$

This formula can be understood by working from the inside out. We write "$C_i(x)$" for the number of violations of constraint C_i incurred by representation x. Given that w_i is the weight of constraint C_i, the expression "$w_i C_i(x)$" stands for the number of violations of a constraint multiplied by the weight of that same constraint. The sum of the weights-times-violations across all of the constraints is written $\sum_{i=1}^{N} w_i C_i(x)$. This sum is negated, "$-\sum_{i=1}^{N} w_i C_i(x)$", and then e (the

[1] The use of constraints is the most widely adopted general approach to phonotactics. The alternative strategy of "licenses", discovered by generalizing over lexical items, is also the subject of current research: for two rather different approaches see Albright (2006) and Heinz (to appear a, to appear b).

base of the natural logarithm) is raised to the result, "$\exp(-\sum_{i=1}^{N} w_i C_i(x))$." We will explain the presence of the minus sign below.

Maxent values have a precise interpretation in the theory of probability: the maxent value of any representation is proportional to its share of the *total probability of all possible representations*. More precisely, if Z is the sum of the maxent values of all representations, then $h(x)/Z$ is the probability of representation x according to the grammar.

In various places, we will refer simply to the sum of the weights-times-violations, without the final negation and exp( ) steps seen in (4). This sum, expressed by the equation $\Phi(x) = \sum_{i=1}^{N} w_i C_i(x)$, will be called the *score* of x according to the grammar.[2]

In example (5), three representations are evaluated by the two constraints *#V ("no word-initial vowel") and *C# ("no word-final consonant"), illustrates scores and maxent values.

(5) *Scores and maxent values for three representations*

x	*#V (w = 3.0)	*C# (w = 2.0)	*Score* ($\Phi(x)$)	*Maxent value* ($h(x)$)
CV	3.0 · 0	2.0 · 0	(3.0 · 0) + (2.0 · 0) = 0.0	exp(–0.0) = 1.00
CVC	3.0 · 0	2.0 · 1	(3.0 · 0) + (2.0 · 1) = 2.0	exp(– 2.0) ≅ 0.14
V	3.0 · 1	2.0 · 0	(3.0 · 1) + (2.0 · 0) = 3.0	exp(– 3.0) ≅ 0.05

Inspection of this table reveals some properties that hold of all of the maxent grammars we propose. Each constraint in the grammar is assigned a weight that is greater than or equal to 0 (i.e., weights are non-negative). Thus a form that is violation-free will receive a score of $\Phi(x) = 0$, hence a maxent value of $\exp(0) = 1$, which is the highest possible value. This is attained in (5) by the candidate CV. Forms with one or more constraint violations have a *lower* maxent value, because maxent value is determined by raising e to the negative of the score: $h(x) = \exp(-\Phi(x))$. The presence of the negative sign can now be understood from the semantics of maxent values: forms with more violations get lower values.

In the learning simulations later in the paper, we will connect theory to data in two ways. When discussing experimental data (§5), we test for a correlation between the experimental observations and the maxent values predicted by a grammar, since the maxent values have a direct theoretical interpretation in terms of probability. On the other hand, when we lack experimental data, as in our study of vowel harmony systems (§6) and stress (§7), it suffices to use just the scores, to establish how well a grammar "separates" well-formed representations from those that are ill-formed. If all well-formed structures receive scores that are lower (i.e. better) than all ill-formed structures, then we judge the grammar to have succeeded in learning

[2] The score, as defined here, is closely related (with a change in sign) to the notion of "harmony" in Harmony Theory (Smolensky 1986).

the non-gradient phonotactic generalizations. This use of scores is equivalent to one in which a language is defined by all and only those representations that surpass a particular threshold.[3]

3.2 Learning maxent weights

Up to this point, we have defined maxent grammars and described how well-formedness is calculated from constraint violations and weights. The main goal of this paper is to develop an algorithm with which such grammars can be learned. It is useful to break the learning problem into two parts: learning the weights of a given set of constraints; and learning the constraints themselves. The first problem has been solved in previous research; the solution, which we review here, explains why the framework is referred to by the name "maximum entropy" and introduces the fundamental comparison between *observed* forms and forms that would be *expected*, or predicted, by a grammar. Previous research has not addressed the second problem in the domain of phonotactics; our proposal is given in the next section of the paper (§4).

The following definition relates the weights of the constraints in a maxent grammar to the data from which the grammar is learned:

(6) *Maxent weighting*

> Given a grammar and a set of observed forms, assign weights to the constraints so that, for each constraint C_i, the expected number of violations of C_i is equal to the number of violations that is observed in the data.

According to a result proven by Della Pietra et al. (1997), assigning weights as in (6) yields a grammar that, subject to the constraints, maximizes the entropy, or disorder, of the distribution of well-formedness values. Intuitively, this means that the predicted well-formedness values must be "maximally noncommittal" (Jaynes 1983) with respect to properties that are not regulated by the constraints. For example, the grammar in (5) penalizes vowels that stand at the beginning of the word (*#V). In the absence of any evidence that particular word-initial vowels are worse than others, this penalty is distributed evenly across all of the vowels in the system. In general, our use of the principle in (6) amounts to the claim that language learners minimize their commitments by assuming uniform (or "flat") distributions in the absence of impinging constraints.

By enforcing the prescription "expected violations = observed violations", one is guaranteed to maximize entropy — and thereby minimize departures from a uniform phonotactic well-formedness distribution that are not motivated by the constraints (and ultimately, by the data from which the constraints are learned). It is a striking fact about maxent grammars that the same prescription also leads to a grammar that maximizes the probability of the observed data (Della Pietra et al. 1997).

The number of times a constraint is violated in a body of learning data (i.e., the observed violation count) is conceptually and computationally straightforward: for each constraint, we

[3] See Hale and Smolensky (2006) for a similar threshold approach.

simply sum the violations of all of the examples in the data. But it may seem odd to talk of the number of violations that are expected, or predicted, by a grammar. How can phonotactic grammars be understood as making such predictions, and how are the expected violations calculated? The general answer makes reference to the set of all possible phonological representations, which we denote by Ω. Formally, an *expectation* over such a set is a weighted average in which the weight of each member of the set is equal to its probability. Therefore, given the general relationship between maxent values and probability (§3.1), the expected number of violations of a constraint is defined as in (7).

(7) *Definition: Expected number of violations*

Given a grammar that determines maxent values, the expected number of violations of constraint C_i is the sum, over all representations $x \in \Omega$, of the probability of x multiplied by the number of violations of C_i that x incurs:

$$E[C_i] = \sum_{x \in \Omega} (1/Z)\, h(x)\, C_i(x)$$

In this equation, "$E[C_i]$" is the expected number of violations of C_i, "$C_i(x)$" is the number of times that representation x violates C_i, and "$h(x)$" is the maxent value of x. The number Z is the sum over all x in Ω of $h(x)$, that is: $Z = \sum_{x \in \Omega} h(x)$. Dividing maxent values by Z transforms them into probabilities.

If Ω were finite (and reasonably small), it would be possible to calculate expected values exactly by enumeration. However, in general there is no upper bound on the length of phonological representations: the set Ω is infinite. And even if there were an upper bound, the number of forms that would have to be enumerated quickly becomes intractable.

Fortunately, there are more sophisticated methods of calculating or approximating expected constraint violations—methods that do not involve enumerating a vast, let alone infinite, set. Both of the methods that we employ incorporate the assumption that the maxent value (hence, probability) of a form approaches zero as the length of the form increases. But they are robust to various assumptions about how this relationship is formalized, and we will not discuss this issue further.

The first method for computing expectations draws upon the same finite-state formalism that has been used in computational OT (Ellison 1994; Albro 1998, 2005; Eisner 1997; Riggle 2004).[4] An individual constraint is stated as a finite-state machine that assigns violations to phonological representations. Constraints are combined into a grammar using intersection (Hopcroft and Ullman 1979). Each path through the resulting machine represents a phonological representation along with its vector of constraint violations. Eisner (2001, 2002), using approximation methods devised by Mohri (2002), shows how to derive the expected values that are needed by summing over all paths through the machine. The summation need not enumerate all of the possible paths, since the sum over the infinite path set is performed in terms of the finite graph.

[4] We are grateful to Jason Eisner for pointing out this connection to us.

The second method for computing the expected number of violations of a constraint is to take a large random sample from the set of all possible representations (Della Pietra et al. 1997). Since the maxent value of a representation can be interpreted as a relative probability, it makes sense to draw a *sample* of phonological representations from the distribution defined by the grammar. When the sample is sufficiently large, the expected number of violations is estimated by the average number of violations in the sample. Generally our samples are of the same order of magnitude as the learning data.

We use the first, finite-state method for determining the weights of constraints that are present in the grammar, and the random-sampling method for the purpose of evaluating the evidence for a new constraint. It would be possible to use the same method for both weighting and constraint evaluation, though at present this would result in either a loss of accuracy or a large increase in computational resources.

We can now describe how these ideas play out in our machine-implemented learner. The process starts with an *a priori* initial weighting; in our simulations, all constraints begin with weight 1.0. The learner also calculates the observed number of violations of each constraint, given the data corpus. The learner then begins an iterative process of reweighting in which first the expected constraint violations are calculated with the current weights, and then the weights are adjusted in order to bring the expected values closer to the observed values. For reweighting, our simulations employ the Conjugate Gradient algorithm (Press et al. 1992). This iterative reweighting will converge to a set of weights that equate expected and observed values.

In sum, our learner employs a variety of well-established algorithms to assign weights to constraints. These weights yield predictions of well-formedness, expressed as maxent values, for phonological forms given their constraint violations.

4. Searching the space of possible constraints

In selecting appropriate constraints, we face the fact that an enormous number of distributional generalizations are consistent with any given set of surface forms. The learner must have a strategy for navigating the space of possible generalizations and selecting members of that space for inclusion in the grammar.

Previous research on phonotactic learning has not addressed the selection problem in a general form. Work in Optimality Theory (Hayes 2004, Prince and Tesar 2004, Jarosz 2006, Pater and Coetzee 2006) generally assumes that the constraint set is provided by UG. No selection problem arises under this approach, as learning consists simply of assigning a ranking to the constraint set. The parameter setting approach set forth by Dresher and Kaye 1990 likewise confronts no selection problem, since the parameters and their cues are provided *a priori*. However, our interest in establishing an inductive baseline (§2.2) is incompatible with any rich UG approach, either constraint-based or parametric. Though it may be necessary to add specific universal constraints to UG, our present goal is to determine how much of phonotactic learning can be done without them.

Another option not open to us is simply to incorporate every constraint into the grammar, relying on the weighting algorithm to determine the importance of each one. This is essentially

the proposal of Pierrehumbert 2006, who applies it to the analysis of medial consonant clusters. This strategy could not be applied in the present context because it does not scale up: the number of constraints grows exponentially as the number of features increases and the domain over which phonotactics are stated expands.

Our proposed solution to the selection problem is twofold. First, we adopt some rather modest UG principles, governing the feature inventory and the format of constraints, which yield a search space that is quite large—hence compatible with the inductive baseline approach—but not intractably so. Second, we employ a set of heuristics to search this space effectively. The next two sections address these two sides of our strategy.

We emphasize that our proposals concerning the search space, like other properties of our learner, constitute a theoretical claim about language learning. To be sure, they are also motivated by issues of implementation—but not, we think, in a way that sacrifices realism with respect to the human learner. If we have characterized the problem of learning phonotactic correctly, then the human learner faces the same search problem as our mechanical learner. The claim is that humans perform the search for phonotactics in a way that is functionally identical to the strategy we describe.

4.1 The constraint space

The learner is assumed to be provided with a set of features, the inventory of segments in the target language, and the feature specifications for each of those segments. In fact, it is the *natural classes* determined by the features, rather than the features themselves, that determine the content of a constraint. Many natural classes have multiple featural definitions, and it is immaterial which particular definition is used to state a constraint. For locating the natural classes determined by a segment inventory and feature set, we use an algorithm and software created by Kie Zuraw.

Using the natural classes, we construct two basic constraint types.

4.1.1 Simple constraints

The first type is just a sequence of feature matrices, as in (8).

(8) $\begin{bmatrix} \alpha F \\ \beta G \\ \vdots \end{bmatrix} \begin{bmatrix} \gamma H \\ \delta I \\ \vdots \end{bmatrix} \dots \begin{bmatrix} \varepsilon J \\ \zeta K \\ \vdots \end{bmatrix}$

where *F*, *G* ... are features and α, β, … take the values + and –. Such a constraint is matched to representations as in *SPE*. It acts as a function, returning the number of matches.

4.1.2 Implicational constraints

We have also found it crucial to incorporate an idea that is central in the literature on phonological constraints (Halle 1959, Stanley 1967, *SPE* ch. 7, Fudge 1969, Prince and

Smolensky 1993/2004), namely *logical implication*; for example "if a particular segment has feature values [αF, βG, …], then any following segment must have the values [γH, δI …]."

An example from the grammar of English onsets is the following: "if a nasal occurs in an onset, any preceding sound must be [s]" (Fudge 1969:279, Selkirk 1982:346). This is straightforward to state as an implication. But without the capacity for implications, we would instead have to formulate a large set of constraints that jointly ban *every segment except* [s] in the context / # ___ [+nas]. Many similar cases can be found.

To formalize implication, we allow exactly one of the matrices of a constraint to be modified by the complementation operator ^; thus [^αF, βG, …] means any segment not a member of the natural class [αF, βG, …]. For example, the limitation of prenasal segments to [s] as discovered by our learner (see (14.17) below) is stated as in (9).[5]

(9) $*\begin{bmatrix} \text{^–voice} \\ \text{+ant} \\ \text{+strid} \end{bmatrix}$ [+nas] "Assess a violation for any segment for any segment which precedes [+nas] and is not [s]."

4.1.3 Limiting the number of possible constraints

The number of possible constraints is proportional to $|C|^n$, where $|C|$ is the number of natural classes and n is the maximum number of feature matrices that may occur in a constraint. $|C|^n$ will be of feasible size if both $|C|$ and n are sufficiently small.

$|C|$ will in general be small to the extent that the feature system makes use of principles of *underspecification*, as embodied in works such as Kiparsky 1982; Archangeli 1984; Steriade 1987, 1995. In keeping with this work, we hypothesize that natural-language phonologies deploy feature systems that use underspecification to achieve relatively low $|C|$ values. In our simulations, we use feature systems embodying both privative underspecification (e.g., [labial], [coronal], and [dorsal] may only take the value +) and contrastive underspecification (e.g., for English [voice] is specified only on obstruents, where it is contrastive).

We have found that feature systems without underspecification cannot feasibly be used in our learner. For example, our feature system for English consonants defines about 200 natural classes, yielding roughly 30 million constraints of three natural classes or shorter. In comparison, the feature system of *SPE*, which makes no use of underspecification, defines more than 600 natural classes; this increases the number of constraints by a factor of about 25, which defeats our software.

Returning to the formula $|C|^n$, we consider ways to limit n. We believe that no particular value of n can be imposed on all constraints, rather that the value of n should be sensitive to the internal complexity of the constraint. In other words, there is a trade-off between the size of a

[5] We will use the following abbreviations for feature names: *ant* = anterior; *appr* = approximant; *cons* = consonantal; *cont* = continuant; *cor* = coronal; *dors* = dorsal; *lab* = labial; *lat* = lateral; *nas* = nasal; *son* = sonorant; *spread* = spread glottis; *strid* = strident; *str* = stress; *syl* = syllabic; *vce* = voice. We will also use C for [–syllabic], V for [+syllabic], # for [–segment] (a word boundary; cf. *SPE*), and [ ] for [+segment] (any segment, also as in *SPE*).

constraint (the number of natural classes that define it) and its specificity. For instance, we suggest that constraints on stress patterns, which manipulate a tiny number of natural classes (defined only by degree of stress and syllable weight), may employ an *n* of up to 4 (§7.3), whereas segmental constraints, which manipulate a far larger set of natural classes, must be limited to size 2, with 3 permitted under special circumstances (§5.1). We postpone the details of our proposals about this trade-off to the discussion of the simulations.

In sum, we limit the constraint space in two ways, reducing the size of *C* with underspecification and of *n* with limits on overall constraint complexity.

4.2 Search heuristics

Underspecification and the size/specificity trade-off ameliorate, but do not solve, the search problem. The set of possible constraints often numbers in the millions even with these restrictions in place. The learner must have a way of homing in on the constraints that are important for characterizing the target language. Our search heuristics take the following form:

(10) a. Search first among the constraints that are most *accurate* (§4.2.1).
b. Among constraints of (roughly) equal accuracy, search for constraints that are maximally *general* (§4.2.2).

4.2.1 Accuracy

We define the accuracy of a constraint as the number of violations of the constraint observed in the data (O), divided by the number of violations expected given the current grammar (E); hence accuracy equals O/E. O is straightforwardly obtainable by applying the constraint to the learning data. E is obtained by generating a random sample of forms with the current grammar (as described in §3.2) and applying the constraint to the sample. Under the reasonable hypothesis that languages favor accurate constraints, one would expect that a constraint with O/E of (say) 0/1000 would be a very powerful constraint whose violation would lead to a strong intuition of ill-formedness, whereas a constraint with O/E of 500/1000 might at best induce a small sense of ill-formedness.

The use of the O/E criterion is justified by empirical work (Pierrehumbert 1994, Frisch 1996, Frisch and Zawaydeh 2001) which shows that it strongly predicts phonotactic well-formedness judgments of native speakers. However, calculation of E has been limited to the special case in which constraints do not overlap. An advantage of the maximum entropy approach is that E can be estimated on a fully principled basis, using the entire set of already-discovered constraints.

Two further refinements are needed. First, we would expect a constraint with O/E of 0/10 to be "weaker" than one with 0/1000, the intuition being that in the first case violations are expected to be rare in any event. To reflect this intuition, we follow the method of adjustment proposed by Mikheev (1997; see also Albright and Hayes 2002, 2003), which substitutes a statistical upper confidence limit on O/E for O/E itself. Using this method, a difference in

accuracy between 0/10 and 0/1000 comes out not as 0 vs. 0, but as 0.22 vs. 0.002.[6] Second, in our implementation we do not actually sort the constraints by accuracy, but rather use an approximate criterion consisting of a stepwise rising accuracy scale (e.g., O/E < .001, OE < .01, and so on).

4.2.2 Generality

Within the strata defined by accuracy, our system selects constraints in order of generality. The idea that the learner of phonology seeks simple generalizations goes back at least to *SPE*, though it has typically been applied at the level of entire grammars or grammar fragments rather than to individual rules or constraints.

We implement generality as a two-level hierarchy. First, *shorter* constraints (fewer matrices) are treated as more general than longer ones. This procedure is effective, because longer sequences can often be assessed on the basis of the shorter sequences they contain. For instance, the well-formedness of a consonant cluster $C_1C_2C_3$ is usually determined by that of C_1C_2 and C_2C_3 (Clements and Keyser 1983, Pierrehumbert 1994). In such cases, early discovery of simple, widely-applicable constraints obviates the need for more complex ones.

From the same principle it follows that among constraints of equal length, we should first search those whose matrices contain the most general featural expressions. The classic way of assessing featural generality is the feature-counting metric of *SPE*. However, in keeping with our overall emphasis on natural classes instead of their featural expressions, we suggest that the value of a constraint is proportional to the number of segments contained in its classes, and we employ a simple metric to sort the constraints on this basis. When subordinated to the feature matrix count criterion, this yields the final search order.

In sum, our learner primarily seeks constraints that are accurate, following an ascending sequence of thresholds for O/E. To chose among constraints at the same threshold, it prefers constraints that are short, and among these, constraints that have more general natural classes. Using these procedures, a constraint space in the tens of millions can be effectively searched. This, in turn, makes our system a fulfillment of our original goal of creating an inductive-baseline learner.

4.3 Learning a phonotactic grammar

The complete process of learning alternates between constraint selection and constraint weighting: a new constraint is selected, following the criteria of §4.2, and then all the constraints are reweighted, following the method of §3.2. This alternating procedure is necessitated by the O/E accuracy criterion for constraint selection. Recall that E is estimated using whatever constraints are already in the grammar. Each newly introduced constraint, once weighted, alters the values for E, and it is these updated values that are relevant for selecting new constraints. Moreover, reweighting must be carried out on the entire constraint set, not just the new

[6] We use a value of $\alpha = 0.975$ for the upper confidence limit, which in our experience largely suffices to exclude pointless constraints from the learned grammars without also excluding constraints with explanatory merit.

constraint, since the new constraint often takes over some of the explanatory burden borne by its predecessors.[7]

The overall algorithm is summarized as in (11).

(11) *Phonotactic learning algorithm*

Input to the algorithm: a set Σ of segments, a set $\mathcal{F}$ of features, a set $\mathcal{D}$ of surface forms, an ascending set $\mathcal{A}$ of accuracy levels, and a maximum constraint size $\mathcal{N}$

1 begin with an empty grammar $\mathcal{G}$
2 **for** each accuracy level a in $\mathcal{A}$
3 **do**
4 select the most general constraint (§4.2.2) with accuracy less than a (if one exists) and add it to $\mathcal{G}$
5 train the weights of the constraints in $\mathcal{G}$ (§3.2)
6 **while** a constraint is selected in step 4

As stated here, the learning algorithm terminates when the search in (11.4) fails to return a new constraint at the least stringent accuracy level. It is also possible, in the interest of expediency, simply to stipulate a maximum grammar size.

In what follows, we first assess the effectiveness of our inductive baseline model against data from a classic area of phonotactic study, the onset inventory of English (§5). We then move away from our inductive baseline, showing the crucial effectiveness of autosegmental tiers (Shona vowel harmony, §6) and the metrical grid (unbounded stress, §7). Our final analysis takes on the phonotactics of an entire language, Wargamay (§8).

5. English Onsets and Gradient Well-formedness

The inventory of syllable onsets in English is an ideal empirical domain for the testing of phonotactic learning models. The basic generalizations have been extensively studied (Bloomfield 1933, Whorf 1940, O'Connor and Trim 1953, Fudge 1969, Selkirk 1982, Clements and Keyser 1983, Hammond 1999), and data are available from experimentation that permit rival models to be evaluated. In this section, we report the results of learning maximum-entropy constraints on word-initial onsets.

5.1 Learning simulation

For our onset simulation, we provided our system with a learning corpus drawn from the word-initial onsets in the online CMU Pronouncing Dictionary.[8] In order to allow the model to

[7] It is for this reason that completed grammars often include constraints of weight zero, functionally equivalent to removing them from the grammar. These usually are constraints that were selected early but rendered obsolete by subsequent discoveries. We omit such constraints in reporting simulation results.

[8] http://www.speech.cs.cmu.edu.

characterize segments as onset-initial or onset-final, we included a boundary element, represented by "#", so that the onset of *string*, for example, was represented as *#str#*.

It is not known to what extent learners of English are affected by exposure to "exotic" onsets such as [zw] (as in *Zwieback*), [sf] *(sphere*), and [pw]) (*Puerto Rico*). To deal with these, we made two onset corpora, one from which the onsets that we intuitively judged exotic were expunged, and another that included all of the onsets that are attested in our own vocabularies. The corpora were created before any modeling was done, so we can claim not to have tailored them to get the intended results. In any event, we found that our own model and others obtained better predictions when trained on the non-exotic corpus, and we report those results here.[9]

The full training data, with frequencies,[10] are given in (12).

(12) *The English onset training set*

k 2764, r 2752, d 2526, s 2215, m 1965, p 1881, b 1544, l 1225, f 1222, h 1153, t 1146, pr 1046, w 780, n 716, v 615, g 537, dʒ 524, st 521, tr 515, kr 387, ʃ 379, gr 331, tʃ 329, br 319, sp 313, fl 290, kl 285, sk 278, j 268, fr 254, pl 238, bl 213, sl 213, dr 211, kw 201, str 183, θ 173, sw 153, gl 131, hw 111, sn 109, skr 93, z 83, sm 82, θr 73, skw 69, tw 55, spr 51, ʃr 40, spl 27, ð 19, dw 17, gw 11, θw 4, skl 1

We used a fairly standard feature set for English, taken mostly from *SPE* and from Halle and Clements (1983). We controlled the total number of natural classes defined (§4.1.3) by using both contrastive and privative underspecification, shown below with blanks.

[9] A more principled alternative would be to model a community of learners, each of which is exposed to a random sample from the larger corpus. Pierrehumbert 2001b shows that this approach sheds light on the relationship between the specificity of a phonotactic constraint and its learnability.

[10] These are type, not token frequencies. Using the latter produces less accurate results in modeling the experimental data discussed below (§5.3). In general, it appears that the use of type frequencies yields better results in modeling any sort of phonological intuitions based on the lexicon; see Bybee (1995, 2001), Pierrehumbert (2001a), Albright (2002), Albright and Hayes (2003), and Hayes and Londe (in press).

(13) *Feature set for English consonants*

	p	t	tʃ	k	b	d	dʒ	g	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	r	j	w
cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	–
appr	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+
son	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+
cont	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+							
nas																		+	+	+				
voice	–	–	–	–	+	+	+	+	–	–	–	–	–	+	+	+	+							
spread													+											
lab	+				+				+					+				+						+
cor		+	+			+	+			+	+	+			+	+	+		+		+	+		
ant		+	–			+	–			+	+	–			+	+	–		+		+	–		
strid		–	+			–	+			–	+	+			–	+	+		–		–	–		
lat																					+			
dors				+				+												+				
high																							+	+
back																							–	+

We set the maximum constraint size *n* (§4.1.3) at 3 and the accuracy schedule (§4.2.1) at (.001, .01, .1, .2, .3). To implement our proposed trade-off between constraint size and featural specificity (§4.1.3), we stipulated that no constraint could contain more than two matrices drawn freely from the full feature set; the remaining matrix of a size 3 constraint was limited to a set of seven "core" natural classes, i.e., the class containing only the boundary marker (#) and the classes [±syllabic], [±consonantal], and [±sonorant].

5.2 The learned grammar

Under these conditions, the learner induced the following 24 constraints in order. Multiple runs yielded identical constraint and weights.

(14) *The learned grammar for English onsets*

Constraint	*Wght*	*Comment*	*Examples*
1. *[+son, + dors]	5.45	*ŋ	*ŋ, *sŋ
2. *[+cont,+voice,–ant]	2.87	*ʒ; see also #22.	*ʒ
3. $*\begin{bmatrix} \text{^–voice} \\ \text{+ant} \\ \text{+strid} \end{bmatrix}$ [–son]	6.83	Obstruents may only be preceded by [s].	*kt, *kk, *skt
4. *[ ][+cont]	6.66	Fricatives may not cluster with preceding C.	*sf, *sθ, *sh, *sfl
5. *[ ][+voice]	6.64	Voiced obstruents may not cluster with preceding C.	*sb, *sd, *sgr
6. $*\begin{bmatrix} \text{^+cons} \\ \text{–son} \end{bmatrix}$ [+cons]	0.86	Obstruents, nasals, and [l] may only be preceded by oral (not [h]) obstruents.	*hl, *hl, *wl, etc.
7. $*\begin{bmatrix} \text{^+cons} \\ \text{–son} \end{bmatrix}$ [+cor]	3.44	Coronals may only be preceded by true obstruents.	*hr, *hl
8. *[+son][ ]	5.36	Sonorants may not cluster with following C.	*jw

9. *[–strid] [^–cons; +son]	3.29	Nonstrident coronals may only be followed by non-[h] glides.	*tl, *dl, *θl, *stl
10. *[ ][+strid]	1.87	Stridents must be initial in a cluster.	*stʃ
11. *[+lab] [^+son; +cor]	5.82	Labials may cluster only with following coronal sonorants.	*pw, *fw, *spw vs. pr, fr, spr
12. * [^–voice; +ant; +strid] [+cons; +lab]	1.30	Labial true consonants may only be preceded by [s].	*ʃm, *km vs. sp, sm
13. *[–ant][^–cons,+cor]	5.45	Nonanterior [coronals] may only cluster with [r].	*ʃl, *ʃw vs. ʃr
14. * [+voice; +cor] [^–cons; +son]	0.56	Voiced coronals may only cluster with [r]	*dl, *zl vs. dr
15. *[–cons][^+back]	2.57	Glides may cluster only with [w].	*hj vs. hw
16. *[+cont,+voice][ ]	5.48	Voiced fricatives cannot cluster with a following C.	*vl, *vr
17. *[[^–voice; +ant; +strid] [+nas]	6.59	Nasals may only be preceded by [s].	*kn, *fn, *spn vs. sn, sm
18. *[+vce,+strid][ ]	1.21	Voiced stridents may not cluster with a following C.	*dʒr
19. *[–cont,–ant][ ]	5.09	Affricates may not cluster with a following C.	*tʃr
20. *[+ant,+strid][–ant]	4.98	Anteriority assimilation	*sr vs. ʃr
21. *[ ][–back]	5.86	[j] may not cluster with a preceding C. (*Beauty,* etc., assumed to have a [ju] diphthong; Clements and Keyser 1983, 42)	*$[bj]_{ons}$
22. *[+cont,+vce,+cor]	2.52	*voiced coronal fricative (violable)	ð, z
23. *[+vce] [^+son; +cor]	3.06	Voiced obstruents may only be followed by coronal sonorants (violable).	gw, dw vs. gr, dr
24. *[–strid] [^–cons; +cor]	3.24	Nonstrident (coronals) may only be followed by coronal glides; i.e. [j] (violable).	tw, θw, stw vs. tr, θr, str

Note that while most of the constraints rule out impossible onsets, the last three learned penalized onsets that are merely strongly underrepresented in the English lexicon.

5.3 Assessing the learned grammar

We assessed the learned grammar first by comparing its predictions with the English lexicon, then by checking it against experiment results.

5.3.1 Comparison with the lexicon

We sought to determine whether the grammar would admit the attested onsets of English (defined as those included in the learning data) and exclude all others. To this end, we created a list of all *logically* possible onsets consisting of up to three English phonemes. Tested on this list, the grammar assigned penalty scores (§3.1) of at least 3.24, to all unattested onsets. In particular, the six lowest scores for clusters not in the learning data were: [stw] 3.24, [sr] 4.98, [tʃr] 5.09,

[jw] 5.36, [ʒ] 5.39, and [ŋ] 5.45. Of these, the least penalized form [stw] has been called an "accidental gap" in the literature (Fudge and Shockey, ms.; Rastle et al. 2002).

Most of the attested onsets received perfect scores (0). However, a few of the rarest onsets in the language did receive penalties: [ð] 2.52, [z] 2.52, [gw] 3.06, [tw] 3.24, [θw] 3.24, [dw] 6.30. Given the rarity of these onsets, only the last score strikes us as being clearly in error.

We conclude that the grammar did a reasonably good job of separating good from bad onsets, the threshold (cf. (§3.1)) falling at a score of about 3.5.

5.3.2 Modeling experimental data

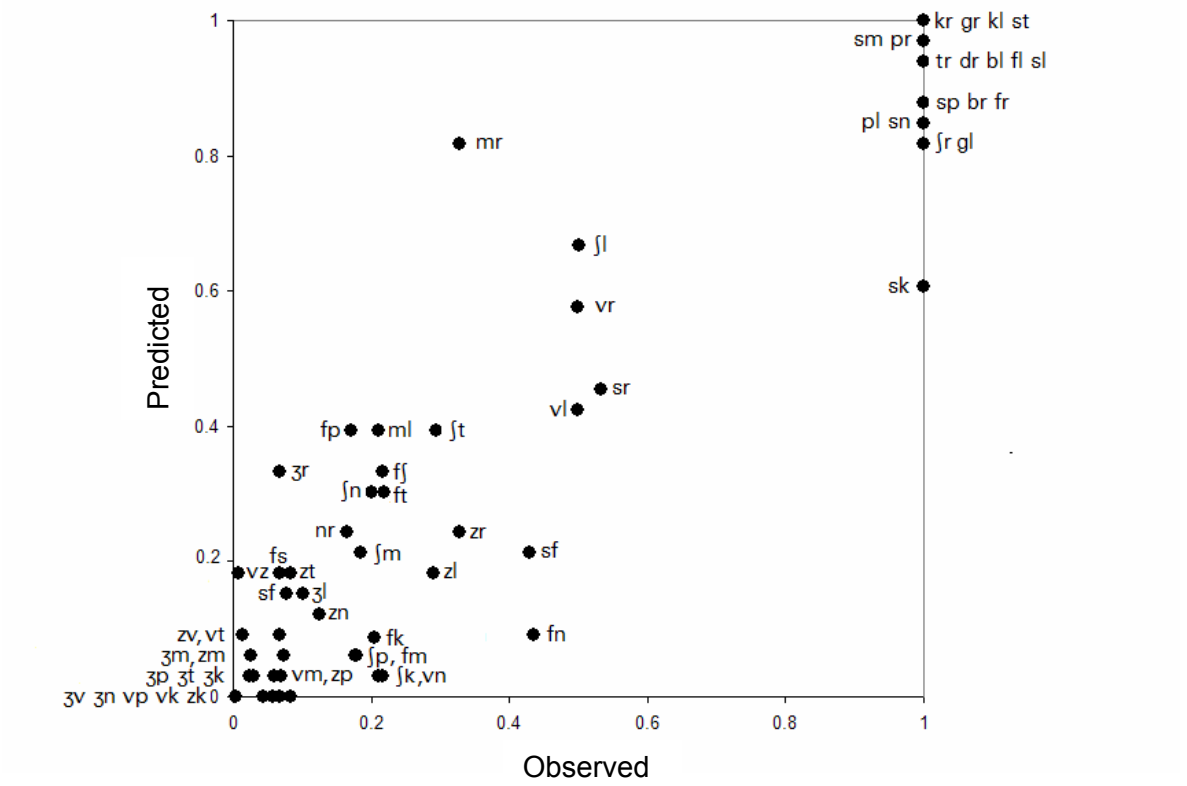
We also assessed the grammar on its ability to replicate the gradient judgments of English speakers. To this end we modeled the data from one of the earliest experiments on phonological well-formedness intuitions, Scholes 1966. Scholes obtained yes/no ratings of 66 monosyllabic non-words from a group of seventh grade students (N=33). The students were asked, for each form, whether it "is likely to be usable as a word of English". The syllable rimes of the nonwords were kept few, and deliberately bland, so that the great bulk of the variation in responses can plausibly be attributed to the onsets. Following a method developed and verified by Coleman and Pierrehumbert 1997 and Frisch et al. 2000, we take the proportion of "yes" responses obtained in Scholes' study, pooled across students, to be an indicator of the mean well-formedness intuition of individuals in the population. Frisch et al. 2000 demonstrate that this method yields scores that are highly correlated with well-formedness ratings on a numerical scale. As with results from similar studies, the pooled Scholes data shows a gradient transition from relatively well-formed to highly ill-formed clusters; see figure (16) below.

As discussed §3.1, our learned grammar assigns a score $-\Phi(x)$ to each possible onset x. In order to render these scores comparable in overall distribution to the experimental data, we introduced one free parameter T, as in (15).

(15) $\text{predicted-rating}(x) = \exp(-\Phi(x)/T)$

Using the best fit value for T, we found that the correlation of predicted ratings against observed ratings (fraction of "yes" responses) was $r = 0.94$. This means that most of the variation in the subjects' responses is explained by the model. The following scattergram shows the predictions of the model plotted against subject ratings for all 62 words in the Scholes experiment.

(16) *Performance of the model in predicting the data of Scholes (1966)*



The correlation of 0.94 becomes more meaningful when compared with the correlations obtained under alternative approaches. We tested five other models, described as follow.

I. In order to compare our machine-learned constraints with a hand-crafted grammar, we translated the constraints proposed by Clements and Keyser (1983) into our own formalism and assigned them weights as in §3.2. [11]

The other four alternatives differ from ours in having substantially less representational power, and were intended primarily to test the usefulness of featural representations in accounting for human well-formedness judgments.

II. The grammar labeled "without features" in (17) below was learned in the same way as our model, but differs in having only single-membered natural classes (one per segment in the inventory).

[11] The constraints consisted of numbers 1, 2, 4, 6, 7, 8, 12, and 15 of (14) above, plus *[–vce][+vce], *[ ][–cont,–ant], *[+cont][+cont], *[+lab][+lab], *[–strid][+lat], *[–vce,+ant,+strid], [–cons,+cor], *[+cont,–vce,–ant][^–cons,+cor], and *[–vce,+ant,+strid][–cont,–vce,+ant][+back].

III. We implemented Coleman and Pierrehumbert's (1997) onset-rhyme model, which has one rule for each onset type in the learning data, and does not use segmental or featural information to relate non-identical onsets.[12]

IV. We also included an *n*-gram model from computational linguistics, constructed with the ATT GRM library (Mohri 2002, Allauzen et al. 2005).[13] This model uses segmental representations, not features, and was trained with standard methods.

V. Lastly, we tested an analogical model patterned after Bailey and Hahn (2001). This model assesses well-formedness not with grammatical constraints, but on the basis of the aggregate resemblance of the word under consideration to all the words in the learning data.[14]

All models were fitted to the data with a free parameter T, just as was done with our own model.

The performance (measured by *r*) of the various models is summarized below.

(17)

Model	*r*
our model	0.94
Clements-Keyser replication (with maxent weights)	0.93
Coleman and Pierrehumbert 1997	0.89
our model without features	0.88
n-gram model	0.88
analogical model	0.83

As can be seen, our machine-learned grammar was sufficiently accurate that it slightly outperformed a carefully hand-crafted grammar. It may be an additional advantage of our grammar that, unlike the Clements-Keyser system, it imposes slight penalties on rare (and to our judgment, slightly marginal) onsets such as [θw]. These two grammars outperform all the others; a plausible reason is they are the only models that employ the standard apparatus of phonological theory, namely features and natural classes. [15]

[12] In training this grammar, we used the standard Good-Turing estimation technique (see, for example, Jurafsky and Martin 2000).

[13] http://www.research.att.com/sw/tools/grm/.

[14] We explored a number of versions of this model and found that the best-performing version was one that used the segmental similarity metric of Frisch, Broe and Pierrehumbert (2004) and that paid no head to token frequencies in the learning data.

[15] A final note concerning these models. Bailey and Hahn (2001) suggest that improvements in modeling can be obtained if constraint-based and analogical models are blended. We find that this is true, but only to a limited extent. If we augment our model with the outputs of the analogical model, *r* rises from 0.94 to 0.95, with comparable levels of improvement in the case of the other models. The relative performance of the models remains unaffected.

6. Nonlocal phonotactics: Shona vowel harmony

The English onset simulation was a demonstration of our model in its simple, inductive baseline version. We consider next a phonotactic pattern that requires us to move beyond the baseline. The pattern in question, involving vowel harmony, is *nonlocal*, imposing restrictions on nonadjacent sounds, since the vowels in question can be separated by strings of consonants.

We focus on the harmony system of Shona, a Bantu language of Zimbabwe (Fortune 1955, Beckman 1997, Riggle 1999.) We chose Shona because it has relatively few exceptions in stems, so that vowel harmony is plainly evident as a phonotactic principle. In this respect Shona differs from other vowel harmony languages (cf. Kiparsky 1973 for Hungarian, Clements and Sezer 1982 for Turkish), where abundant disharmonic stems would create problems for a purely phonotactic learning strategy. For the same reason, we limit our study to verbs, where the harmony pattern is closest to exceptionless.[16]

6.1 The Shona data pattern

Shona has five vowels: [i e a o u], whose distribution is restricted by the harmony principles given below (examples, given in Shona orthography, are from Hannan 1981):

(18) *Shona vowel distribution*

a. *e*, *o* may occur as follows:
   i. in initial syllables, as in *beka* 'belch', *gondwa* 'become replete with water'.
   ii. *e* may occur non-initially if the preceding vowel is *e* or *o*, as in *cherenga* 'scratch', *fovedza* 'dent'.
   iii. *o* may occur non-initially if the preceding vowel is *o*, as in *dokonya* 'be very talkative'.

b. *i*, *u* may occur as follows.
   i. in initial syllables, as in *gwisha* 'take away', *huna* 'search intently'.
   ii. *i* may occur non-initially unless the preceding vowel is *e* or *o*, as in *kabida* 'lap (liquid)', *bhigidza* 'hit with thrown object', *churidza* 'plunge, dip'.
   iii. *u* may occur non-initially unless the preceding vowel is *o*, as in *baduka* 'split', *bikura* 'snatch and carry away', *chevhura* 'cut deeply with sharp instrument', *dhuguka* 'cook for a long time'.

c. *a* is freely distributed.[17]

In dynamic terms, this implies a kind of asymmetrical harmony for [high]: the mid vowels *e*, *o* require a following high *i* to be lowered to *e*, and the mid vowel *o* requires a following *u* to be lowered to *o*. In fact, Shona suffixes alternate in height in order to remain in conformity with these requirements (Fortune 1955:26, Beckman 1997:10-11), though our focus is on harmony as a phonotactic pattern.

---

[16] For the idea that particular parts of speech have special phonotactics, see Smith (2001).

[17] However, in our learning data, final vowels are always /a/, since the dictionary entries for verbs all end with the suffix /-a/.

We analyzed 4399 Shona verbs taken from the online version of Hannan's (1959) Shona dictionary, available from the CBOLD project.[18] Inspection of the corpus showed that even in verbs, the harmony system is not free of exceptions: a fair number of idiophones and borrowings violate the normal harmony pattern. The details are given in chart (18), which gives totals from our training set for all 25 possible two-vowel sequences. The chart gives both the raw counts and an *ad hoc* O/E estimate, namely the raw frequency divided by the product of the two individual vowel frequencies. The latter gives a clearer notion of over- or under-representation by compensating for the overall frequencies of vowels. Phonotactically aberrant cases are classified intuitively as "✓", "?", or "*" according to the kind of violation they contain.

(19) *Shona vowel distribution: corpus data*

Vowel sequence	*Count*	*Ad hoc O/E*	*Status*	*Classification*
a a	1539	0.63	✓	
a e	**3**	**0.006**	*	Noninitial *e* without harmony trigger
a i	532	0.83	✓	
a o	**1**	**0.002**	*	Noninitial *o* without harmony trigger
a u	587	0.627	✓	
e a	655	1.28	✓	
e e	601	5.65	✓	
e i	**3**	**0.023**	*	*i* not lowered after *e*
e o	**0**	**0.00**	*	Noninitial *o* without harmony trigger
e u	264	1.33	✓	
i a	1220	1.91	✓	
i e	**0**	**0.00**	*	Noninitial *e* without harmony trigger
i i	507	3.05	✓	
i o	**0**	**0.00**	*	Noninitial *o* without harmony trigger
i u	179	0.72	✓	
o a	650	1.25	✓	
o e	156	1.44	✓	
o i	**23**	**0.17**	?	*i* not lowered after *o* (weak trigger)
o o	705	6.37	✓	
o u	**20**	**0.10**	?	*u* not lowered after *o* (weak trigger)
u a	1815	1.91	✓	
u e	**4**	**0.02**	*	Noninitial *e* without harmony trigger
u i	180	0.72	✓	
u o	**1**	**0.005**	*	Noninitial *o* without harmony trigger
u u	845	2.28	✓	

The detailed data illustrate an aspect of Shona that has to our knowledge not been previously noticed: *o* is somewhat "weak" as a harmony trigger, in that the high vowels *i* and *u* follow it with modest frequency. The sequences *o i* and *o u* are nevertheless underrepresented, and we will assume that a phonotactic grammar should take account of this; this claim is reflected in our assignment of "?" status to these sequences.

[18] http://www.cbold.ddl.ish-lyon.cnrs.fr/.

6.2 Failure of the inductive baseline model

Adopting a straightforward feature system for Shona segments, we ran our inductive baseline learner on Shona. The settings were the same as for the English onset simulations (§5.1), with a few exceptions. We set *n* (the maximum number of feature matrices in a constraint) at 4, the reason being that for V_1CCV_2 sequences, the harmonic dependency between V_1 and V_2 has no chance of being detected unless constraints can span four feature matrices. In addition, we extended the highest accuracy threshold (§4.2.1) very slightly from 0.3 to 0.35, which (ultimately) proved necessary for capturing the marginal status of *o u* and *o i*. We ran the learner more or less at the limit of our equipment (a multi-day run), gathering 200 constraints.

To test the resulting grammar, we gave it 50 test words to rate. Of these, 25 took the form *m***V***m***V***ma*, where the two slots labeled "**V**" were filled with all possible vowel pairs (*mimima*, *mimema*, etc.) The remaining 25 were similar, but took the form *m***V***nd***V***ma*, chosen to test if the system had learned the harmonic restrictions across consonant clusters.

The inductive baseline model achieved only minimal descriptive success. It did find five valid harmony constraints applicable to VCV sequences, shown in (20).

(20) *Results of the inductive baseline learner applied to Shona*

	Constraint	*Weight*
a.	**a* [ ][–high, –low]	4.75
b.	*[–back][ ] *o*	4.70
c.	*[+high][ ][–high, –low]	4.22
d.	*[–high, –low][ ] *i*	3.51
e.	**a* [ ] *o*	1.80

These sufficed to rule out all the ill-formed cases of *m*V*m*V*ma*; they also penalized ?*momima* (with the same value as **memima*), and left only ?*momuma* classified erroneously as perfect. However, for the *m*V*nd*V*ma* forms, the model failed completely: no constraints regulating the vowels of V_1CCV_2 were found, so all of these were classified as perfect, irrespective of their vowels.

We judge that the reason for this failure lay in the unmanageable hypothesis space. The number of four- matrix constraints is very large, and the available search time was consumed before the relevant V-to-V constraints could be found.[19] The presence of a not inconsiderable number of *triple* clusters (e.g. [ndw]) in Shona renders the possibility of the inductive baseline learner succeeding even more remote.

[19] As a control, we also considered whether the Shona lexicon fails to instantiate the principles of harmony properly for vowel pairs that are separated by consonant clusters. To test this, we made up a set of pseudowords, of the form VCCV, extracted from all such sequences in the real training set (e.g., [iŋga] from *tʃiŋgamidza*. Our improved vowel-projection learner (see below, §6.3) learned from these a reasonable approximation of the harmony pattern. This shows that the failure of the inductive baseline learner cannot be attributed to gaps in the learning data, but rather must be the result of the wrong learning strategy.

6.3 Moving beyond the inductive baseline: projections

From the viewpoint of contemporary phonological theory, the analytic approach to vowel harmony offered by our inductive baseline system is a rather unlikely one. Phonologists have long been aware that vowel harmony systems normally "care" only about the vowels of the string, and have adopted formal devices that permit this, expressing the nonlocal process in local terms. This can be done, for instance, with an autosegmental tier for vowels (Clements 1976, Goldsmith 1979), perhaps incorporated into some conception of feature geometry (Archangeli and Pulleyblank 1987, Clements and Hume 1995). Without attempting to choose between these theories, we argue that a vocalic representation can make certain harmony systems learnable that would not otherwise be.

To create the effects of a vowel tier within the computational limits of our system, we use a slightly different conception, due originally to Vergnaud (1977; see also McCarthy 1979). We assume that every phonological representation automatically generates a *vowel projection*, which is a substring consisting of all and only its vowels, appearing in the same order as in the main representation. Projections are scanned during the discovery of phonotactic constraints in the same way as the full representation, and that every phonotactic constraint applies on its own projection. Thus, for example, the constraint *[–high, –low][+high], as defined on the vowel projection, forbids mid-high vowel sequences irrespective of how many consonants intervene.

A projection is defined by a set of criterial feature values, and consists of feature matrices containing only the values of the projected features. For example, the vowel projection employs the single criterial feature value [+syllabic], and projects the features that classify vowels, which for our Shona feature set are [high], [low], and [back]. We assume that projections also include the *SPE* feature [segment], which (in its minus value) designates the word boundary. To give an example, the verb *gondwa*, from (18), is shown in (21) in both its complete representation (which we will call the "default projection") and its vowel projection.

(21)

[–seg]	g	o	n	d	w	a	[–seg]	*Default projection*
[–seg]	[+high, –low, +back, +seg]					[–high, +low, +back, +seg]	[–seg]	*Vowel projection*

We amplified our inductive baseline learner to create projections and scan them for phonotactic generalizations. The modified version of the learner alternates among the available projections, learning from each in turn. For the purpose of determining maxent values (§3.1), the constraints on all projections apply in parallel.

6.3.1 A vowel projection-based grammar for Shona

With this augmented capacity in place, we reran the Shona simulation. The learner quickly found all available constraints on the vowel projection; they were always among the first 30 learned. Since the constraints from the default projection are of little interest here, we therefore

allowed learning to terminate after 30 constraints. We ran the model ten times, and give a representative result in (22). [20]

(22) *Grammar learned for Shona: vowel projection constraints* [21]

	Constraint	*Weight*	*Comment*
1.	**a* [–high, –low]	5.81	*e* and *o* are unlicensed after *a*.
2.	*[+high][–high, –low]	5.46	*e* and *o* are unlicensed after high vowels.
3.	*[–low, –back] *o*	5.01	Neither *i* nor *e* triggers mid harmony, so *o* after these vowels is unlicensed.
4.	**o* V *u*	4.19	Long-distance harmony
5.	*[–high, –low] *i*	2.56	See discussion below.
6.	**o u*	2.16	*o* is a relatively weak harmony trigger.
7.	**e i*	1.98	With #5, makes *e* a strong harmony trigger.

The constraints of the grammar have straightforward interpretations in terms of the generalizations given earlier, with the exception of the last three constraints. #5 and #6, with relatively low weights, combine to make *o* a weak harmony trigger, resulting in the lower penalties for *o i* and *o u* seen in (23) below. Words containing *e i* violate both #5 and #7, receiving a combined score of 4.54. Thus, the grammar predicts that *e* should be a strong harmony trigger for front vowels.[22]

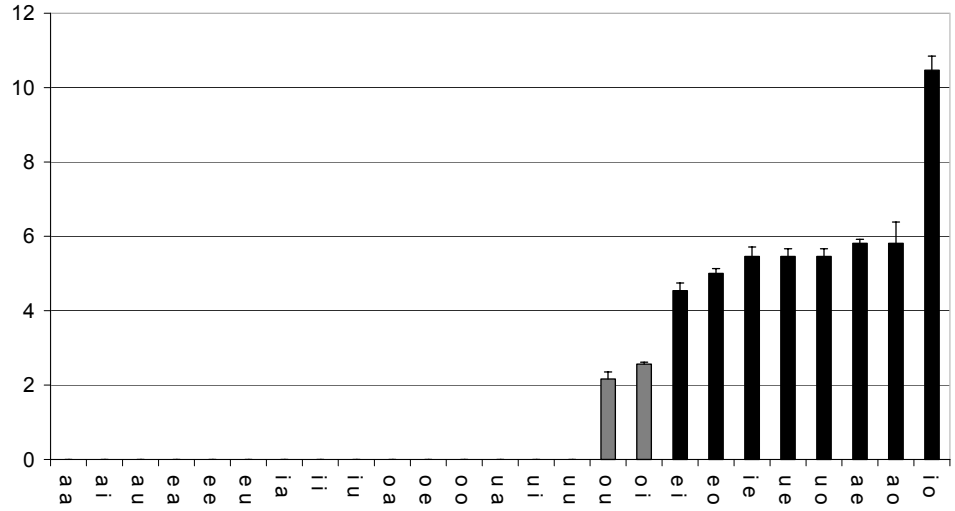
We tested the grammars learned by giving them the same test words to rate (*mVmVma*, *mVndVma*) on which the inductive-baseline grammar had failed. The results are shown in Fig. (23). Since corresponding pairs with these two frames (e.g. *mimima*, *mindima*) emerged with identical predictions, they are conflated in the figure.

[20] Of the ten test runs, six yielded exactly the same vowel tier constraints; of these, we present results from the grammar whose test scores best correlated with the mean (across 25 vowel patterns). The worst pairwise correlation for any two learning runs was $r = 0.984$. The variation in empirical predictions, which are small, can be seen in the 1 s.d. error bars in Fig. (23) below.

[21] One constraint relevant to vowel sequencing was learned on the default tier: a requirement that all final vowels must be [a]. We omit this constraint from (22), since it merely expresses an arbitrary property of the training set, namely the use of [a]-final citation forms (fn. 17).

[22] Two outlier grammars in our series of ten simulations avoided these complications, learning the more intuitive constraint **o* [+high].

(23) *Scores for Shona vowel sequences from projection-based learner.*



We claim that these predictions represent a good approximation of Shona vowel phonotactics. All forms marked as "✓" in (19) above received perfect scores (0). All forms that are fully illegal under the principles of Shona vowel harmony ("*" in (19)) received substantial penalty scores (black bars). Finally, the intermediate forms where *o* does not trigger lowering ("?" in (19)) receive consistently intermediate scores, shown in dark gray.

The system further predicts that of the bad forms, **mimoma* and **mindoma*, which have a vowel sequence nowhere seen in the training set, should be worse than the others. To verify this prediction (essentially that both "*" and "**" forms occur), experimental data on Shona would be required.

In conclusion, we claim that that our learner has achieved a reasonable approximation to Shona vowel sequencing phonotactics. The learning of Shona harmony became possible only when we moved beyond our inductive baseline model to incorporate a vowel projection. Thus, the concept of a vowel tier (or close equivalent) can be defended on learnability grounds: in controlled comparative simulations, it proves essential to phonotactic learning.

7. Locality in stress patterns: the metrical grid

Another type of non-local phonotactics is found in stress systems. Where stress is predictable, it is often analyzed derivationally: a grammar assigns a stress contour to each form, based on its segmental or syllabic representation. But predictable stress is also a phonotactic pattern, a regularity of surface forms. We adopt this perspective here, noting that it readily extends (Selkirk 1980b) to languages like English where stress is not fully predictable, but obeys important restrictions.

7.1 A simple case: unbounded stress

The locality of stress is seen clearly in so-called unbounded stress patterns. One such pattern, attributed to Eastern Cheremis and various other languages (Hayes 1995:§7.2), works as follows:

(24) a. Every heavy syllable has some degree of stress.
b. Every initial syllable bears some degree of stress.
c. Of the stressed syllables in a word, the rightmost bears main stress.

In terms of main stress, this is a "default to opposite" system (Prince 1985), with the pattern "rightmost heavy, else initial." The generalization in (24c) has been called "End Rule Right" (Prince 1983), a term we will use below.

This description implies two clear instances of nonlocality. First, the fact that exactly one syllable bears main stress ("culminativity"; Hayes 1995) is a nonlocal observation that cannot be stated in our inductive baseline model. Because our model imposes a limit *n* on the number of feature matrices that may appear in a constraint (§4.1.3), it can only require that the main stress appear within *n* syllables of a particular word edge or other landmark; it cannot quantify over all syllables of a word to determine that exactly one of them bears main stress. The same considerations imply that the End Rule Right restriction in (24c) cannot be captured by our inductive baseline learner; there is no guarantee that this syllable will fall within the *n*-syllable limit.

A representation for stress that addresses these locality issues is the *metrical grid*, proposed by Liberman (1975) and employed as the basis of general theorizing on stress patterns by Liberman and Prince (1977), Prince (1983) and much subsequent work. From the earliest work, theorists have recognized the implications of the grid for phonological locality.[23]

In a typical grid, every syllable is assigned to a terminal level grid position, represented here as a row of *x*'s at the bottom of the grid. Every stressed syllable is designated as such by assigning it an additional *x* on the second row up from the bottom. Main stressed syllables are also assigned an *x* on the third, highest row. This is illustrated in the following representation, which shows a schematic 13-syllable word containing three heavy syllables, obeying the stress pattern in (25).

(25)

								x						*Main stress row*
x			x		x			x						*Stress row*
x	x	x	x	x	x	x	x	x	x	x	x	x		*Syllable row*
L	L	L	H	L	H	L	L	H	L	L	L	L	word	*Syllables: H = heavy, L = Light*

Since all of this phonological material belongs to a single word, the brackets for word division enclose all levels of the grid.

Grid formalism makes it possible to characterize the nonlocal patterns described above in local terms. The requirement that every word have exactly one stress is expressed by requiring

[23] It has often been argued that the grid should be amplified with constituency information, such as foot structure (Liberman and Prince 1977, Halle and Vergnaud 1987, Hayes 1995). The present discussion makes no use of such constituency, taking an agnostic view on whether it exists. For discussion of "hidden structure" of this kind, see §9.2.

that there be exactly one *x* between each pair of word brackets on the main stress row. The constraint for the End Rule Right generalization can be expressed locally as in (26).

(26) * x *Main stress row*
x x *Stress row*

The literal interpretation is, "avoid a main stress mark when another grid mark follows on the immediately lower row."

7.2 Formalizing grids as projections

The idea, then, is that by providing localist representations of patterns that would appear as non-local in an inductive-baseline representation, the grid should make possible the learning of stress generalizations that would otherwise be missed. To test this idea, we constructed a formalization of the grid, using the same device of projection used earlier for vowel harmony.

For simplicity, we assumed an inventory of terminal elements consisting of just six symbols, each designating a syllable type: { L̆, ˌL, ˈL, H̆, ˌH, ˈH }. L designates light syllables and H heavy; and the IPA diacritics [˘ ˌ ˈ] designate stressless, secondary stressed, and main stressed syllables. These six entities were classified with a simple system of prosodic features: [heavy], [stress], and [main]; where primary stress is [+stress, +main] and secondary stress is [+stress, –main]. We used these features (plus the *SPE* feature [segment], which distinguishes segments from word boundaries) to express the grid as a set of projections:

(27) *Formalizing a metrical grid with projections*

	Selecting features	*Projected features*
a. Main projection	[+main]	[segment]
b. Stress projection	[+stress]	[segment], [main]
c. Default projection	*none*	*all*

The three projections are shown in detail for a schematic form in (28).

(28) *Forming a grid with projections: representation for* [ˌL L̆ ˈH L̆]

[–seg]			[+seg]		[–seg]	*Main projection*
[–seg]	[+seg, –main]		[+seg, +main]		[–seg]	*Stress projection*
[–seg]	[+seg, –heavy, +stress, –main]	[+seg, –heavy, –stress, –main]	[+seg, +heavy, +stress, +main]	[+seg, –heavy, –stress, –main]	[–seg]	*Default projection*

This representation is closely analogous to a traditional grid, we can be seen if one simply replaces every matrix containing [+segment] with *x* and mark word boundaries with brackets, as shown in (29).

(29)

[			x		*Main stress row*
	x		x		*Stress row*
	x	x	x	x	*Syllable row*
	L	L	H	L	]word

The projection version may appear to be richer in information, since each row encodes the presence of higher level grid marks with its featural content. However, traditional use of grids has generally done more or less the same, relying on geometrical ("dominated by") rather than featural descriptions. We judge that our version, which is computationally tractable, represent a reasonable approximation to the original intent of grid theory.

7.3 Learning stress with grids

We tested our learner under this scheme by having it try to learn the schematic stress pattern in (24). As training data, we employed all concatenations of length five or fewer of the symbol set { L̆ ˌL ˈL H̆ ˌH ˈH }that obey (24) ([ˈL], [ˈH], [ˈL L̆], [ˌL ˈH], [ˈH L̆], [ˌH ˈH], [ˈL L̆ L̆], and so forth). With this training set, the projections of (27), and the same learning parameters as in the English onset simulation (§5.1), our system discovered the five constraints in (30).[24]

(30) *Grammar learned for stress pattern (24)*

	Constraint	*Projection*	*Weight*	*Comment*
1.	*# #	Main	6.68	Culminativity
2.	*[+main][+stress]	Stress	9.86	End Rule Right
3.	*H̆	Default	7.33	WEIGHT-TO-STRESS (Prince and Smolensky 1993:56)
4.	*# [–stress]	Default	10.46	Every word must begin with a stress.
5.	*[ ][+stress, –heavy]	Default	9.34	Light syllables may be stressed only if initial.

We tested this grammar by calculating the scores it derives for every possible string up to length five drawn from the complete inventory { L̆, ˌL, ˈL, H̆, ˌH, ˈH }. The grammar successfully assigned perfect scores to all legal forms, and penalty scores of at least 6.68 to all illegal ones.

Moreover, our inductive baseline learner cannot learn this stress pattern. Indeed, it cannot even represent the grammar that would be needed: if the maximum number of matrices used in a constraint is *n*, the grammar will be defeated by words of length *n* + 1. Thus, when we set *n* at 4,

[24] Multiple runs yielded identical constraints and weights.

the grammar learned failed to rule out five-syllable forms like *[ˈH L̆ L̆ L̆ ˈH], with two primary stresses and *[ˌH L̆ L̆ L̆ L̆] (with none).

In sum, hierarchical representations permit the statement of nonlocal generalizations using formal principles that are stated locally. In previous work, this property has been noted as an important basis for developing a constrained theory of possible stress patterns (see, e.g., Hayes 1995:34). But by the same token, the locality property is important to learning, since it makes it possible to discover the crucial generalizations using a learner with a sharply restricted search space.

7.4 Other stress rules

To get a clearer idea of the performance of the model in learning stress systems, we let it attempt to learn similar schematic simulations for the empirical typology of quantity-insensitive systems compiled by Gordon (2002). Gordon's research interest was in developing an *a priori* constraint set whose factorial typology (Prince and Smolensky 1993:§3.1) would match with the observed natural language systems. Here, we simply use his 33 observed stress patterns as a criterion for our model, to determine whether they could all be learned.

Our simulations were done along the same lines as in §7.2, except that since the languages in question make no distinction of syllable quantity, the terminal vocabulary was limited to just three elements distinguished by stress level (σ̆, ˌσ, ˈσ). We followed Gordon in including words of up to eight syllables in the training sets, and in a few cases made minor corrections to Gordon's typology, making use of the cited source materials.

For *n* (the maximum number of matrices in a constraint), we employed a value of 4. This follows our earlier claim (§4.1.3) that constraint systems permit a trade-off of length against internal complexity. Since the feature system for prosodic properties (here, just {[±stress], [±main]}) is impoverished, a value of 4 is feasible without creating a huge search space. Setting *n* at 4 permits the system to learn constraints like [–main][ ][ ] #, which is used for deriving antepenultimate stress (see, for example, the entry for Georgian in Appendix A).

The 33 grammars learned by our system contained a variety of constraints, of which the six most common are given in (31).

(31) *Commonly learned stress constraints*

	Constraint	*Tier*	*Languages/33*	*Comment*
1.	* # #	Main	33	Culminativity
2.	*[+stress][+stress]	Default	21	*CLASH (Prince 1983)
3.	*[+main][+stress]	Stress	14	End Rule Right
4.	*[+stress][+main]	Stress	13	End Rule Left
5.	*[–stress][–stress]	Default	12	*LAPSE (Prince 1983)
6.	*[ ][+stress] #	Default	12	See §9.1.

We tested the 33 learned grammars by examining all possible strings of length 8 or less composed of the elements (σ̆, ˌσ, ˈσ). This test showed that our model was entirely successful in distinguishing the well-formed from the ill-formed strings, assigning a perfect score to every legal form and a substantial penalty to every illegal one, in each language. For the full set of learning data and grammars, see Appendix A. The constraints are discussed further in §9.1.

8. A whole-language analysis: Wargamay

The ultimate goal of our learning model is to induce a complete description of the phonotactics of any given language. In this section, we take a first step toward this goal by applying the model to data from the Australian aboriginal language Wargamay (Dixon 1981). Wargamay was chosen because of its interesting quantity-sensitive stress system, and because Dixon's meticulous description of its phonotactics provides a baseline against which our learned grammar can be evaluated (see also Sherer 1994, Hayes 1995, Kager 1995, McGarrity 2002). The theoretical issues addressed here are similar to those discussed earlier. In particular, our study of Wargamay provides further evidence for the utility of multiple projections in phonological representations, and reveals gradient well-formedness patterns that are not fully accounted for by previous work on the language.

8.1 Segments, features, and training data

The phoneme inventory of Wargamay is given below in IPA.

(32) *Wargamay phonemes*

Consonants		Labial	Apico-alveolar	Retro-flex	Lamino-palatal	Velar
Stops		b	d		ɟ	g
Nasals		m	n		ɲ	ŋ
Trill			r			
Approximants	lateral		l			
	central	w		ɻ	j	

Vowels	Front	Central	Back
high	i, iː		u, uː
low		a, aː	

These phonemes have various allophones, involving contextual or free variation, as well as optional neutralizations, as described in Dixon (1981:16-17). We idealize somewhat in abstracting away from these details.

The features we assumed are as in (33).

(33) *Feature chart for Wargamay*

	b	d	j	g	m	n	ɲ	ŋ	r	l	ɻ	j	w	ĭ	ă	ŭ	ˌi	ˌa	ˌu	ˈi	ˈa	ˈu	ˈiː	ˈaː	ˈuː
syl	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+
cons	+	+	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–
appr	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
son	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lab	+				+																				
cor		+	+			+	+		+	+	+														
ant		+	–			+	–		+	+	–														
lat									–	+	–														
dors				+				+																	
high												+	+	+	–	+	+	–	+	+	–	+	+	–	+
back												–	+	–	+	+	–	+	+	–	+	+	–	+	+
long														–	–	–	–	–	–	–	–	–	+	+	+
stress														–	–	–	+	+	+	+	+	+	+	+	+
main														–	–	–	–	–	–	+	+	+	+	+	+

We used as our learning data the vocabulary of approximately 950 items included in Dixon's grammar. We removed reduplicated forms, which Dixon treats as two separate phonological words, and a handful of forms that contain blatant violations of the phonotactic system.[25] The set of remaining forms was considerably smaller than the learning data for our previous analyses and contained only one item of more than four syllables. We judged this corpus to be too limited to serve as input to our learner, particularly because we are interested in learning the stress system of the language. Therefore, we inflected each nominal and verbal root according to the morphological description given by Dixon (1981:27ff.). The resulting set contains about 6000 words and instantiates the stress pattern across a range of word lengths from one to six syllables. In the following sections, we discuss how the grammar learned from these data accounts for the segmental and stress phonotactics of Wargamay.

8.2 Learning simulation

The resources our learner used for Wargamay integrate those from the previous sections. We deployed projections for a metrical grid (as in (28)) as well as a vowel projection (§6.3).[26] Our Wargamay grids were amplified versions of what was used in the previous section, since instead of just schematic syllable strings, we had to deal with complete representations. Sidestepping the question of syllabification (see §9.2 below for discussion), we defined a Weight projection

[25] These are [gilaɲgan] 'old woman' ([ɲg] cluster, derived from [gilaɲ] 'old man'), [ŋijaɲma] 'ask' ([ɲm] cluster), [jawuɟɲbaɻi] 'big kangaroo' ([ɳɲb] cluster, with the phonotactically-regular variant [jawujmbaɻi]), and the loanwords [drajga] 'tracker' (initial cluster), [gaːguɻuɟ] 'cockroach' (final obstruent), and [lajn] 'line' (initial [l] and final cluster).

[26] As it happened, no constraints were learned on the Vowel projection. We included it for the sake of realism; language learners are not generally told in advance whether their language will have vowel harmony.

whose selecting feature was [+syllabic], and whose projected features were [long], [stress], [main], and [segment]. Since only long-voweled syllables count as heavy in Wargamay, this sufficed to provide a lowest-level grid layer that could represent syllable count and weight.

The feature matrix limit *n* was set at 4 for the grid projections (see §7.4) and 3 elsewhere; otherwise, all parameter settings for the learner were set the same as in the English onset simulation (§5.1).

The system learned a large grammar, which we limited by fiat to 100 constraints. Multiple runs yielded essentially identical results, and we discuss only one representative run here. In what follows, rather than covering the whole grammar all at once, we will divide Wargamay phonotactics into various empirical domains, discussing the constraints learned and the system's performance for each.

8.3 Segmental phonotactics

8.3.1 CV sequencing

The sequencing of consonants and vowels in Wargamay phonotactics is straightforward. Every word must begin with a consonant, vowel sequences are not permitted, and consonant clusters cannot appear at the beginning or end of the word. The grammar constructed by our learner accounts for these restrictions economically:

(34) *Constraints on CV sequencing*

Constraint	*Weight*	*Comment*
*#V	3.64	No initial vowels
*VV	5.84	No vowel sequences
*#[ ]C	5.73	No initial consonant clusters (given *#V)
*CC#	2.92	No final consonant clusters

8.3.2 Initial and final consonants

Any consonant except [r] or [l] (the anterior approximants) can appear at the beginning of the word, and a subset of the sonorants ([m n ɲ l r j]) may appear word-finally. The learned grammar captures these restrictions with the following constraints:

(35) *Constraints on word-initial and word-final consonants*

Constraint	*Weight*	*Comment*
*# [+appr,+ant]	4.93	No initial [r] or [l]
*[–lat] #	4.97	No final [r] or [ɻ]
*[–son] #	4.59	No final obstruents
*[–syl,+back] #	3.84	No final [w]
*[+dors] #	3.79	No final dorsals
*[+lab] #	3.63	No final labials
*[+appr, –ant] #	1.95	No final [ɻ]

With these constraints, the grammar penalizes all of the unattested word-initial and word-final consonants, and gives all of the attested word-initial consonants perfect scores. However, with respect to word-final position it is more restrictive than Dixon's description. The constraint *[+lab]# penalizes both [b], which is not possible finally, and [m], which does occur in that position. Similarly, the constraint *[–lat]#, whose weight seems too high to us, penalizes both unattested [ɻ]# and attested [r]#.

Inspection of the learning data explains why the learner singles out [m] and [r], among the consonants that are attested finally, as relatively ill-formed. There are exactly four items in Dixon's vocabulary that end in [m], and none of the inflectional affixes end in this consonant. Consequently, [m]-final words make up less than 1% of the consonant-final words in the inflected learning data. The numbers for [r]-final words are only slightly higher (11 vocabulary items, < 1% of consonant-final learning data). In comparison, all of the other attested word-final consonants occur in at least 5% of the consonant-final inflected words of the learning data. Thus the learner has selected constraints against the attested word-final consonants that are substantially rarer than their competitors.

8.3.3 Intervocalic consonants and clusters

The richest area of Wargamay's segmental phonotactics is its inventory of intervocalic consonant sequences. Every single consonant is attested intervocalically, and our learned grammar contains no constraints against consonants in the environment / V ___ V. However, the inventory of biconsonantal and triconsonantal clusters is quite restricted. Dixon identifies the following cluster types as possible root-internally:

(36) *Legal root-internal consonant clusters in Wargamay (Dixon 1981)*

a. homorganic nasal-stop sequences ([mb nd ɲɟ ŋg])
b. [n l r ɻ j] followed by [b ɟ g m ɲ ŋ] (i.e., by the set of non-apical stops and nasals)
c. [l r ɻ j] followed by legal nasal-stop sequences or [w]

Many of the clusters included in this description are unattested ([nɲ lɲ ɻm ɻɲ jɲ lnb lnd lnɟ lng rmb rnd rnb rnɟ rng ɻnb ɻnd ɻnɟ ɻng jmb jnd jɲɟ jng jw]) or occur in just one vocabulary item ([nŋ jŋ rw ɻɲɟ ɻŋg ɻw jnb). A generalization not noted by Dixon is that triconsonantal clusters containing *non-homorganic* nasal-stop sequences are marginal: [lnb lnɟ lng rnb rnɟ rng ɻnb ɻnɟ ɻng jng] are among the unattested clusters, and [jnb jnɟ] occur in only one or two roots. A generalization evidently related to (36b) that the apical sequence [nd] never occurs within a triconsonantal cluster (thus [lnd rnd ɻnd jnd] are unattested in roots).

Because roots can end in consonants, and some of the inflectional affixes are consonant-initial, one might expect a much larger inventory of intervocalic clusters in conjugated forms (Dixon 22). To a large extent, this expectation is dashed by morpheme-specific alternation. For example, the ergative/instrumental case ending, which is [-ŋgu] after vowels, loses its initial nasal and undergoes place of articulation assimilation when combined with nasal-final roots; this alternation, like others documented by Dixon, serves to reinforce the phonotactic pattern found root-internally. There are, however, clusters that appear only under inflection, namely [mg mɲ nɲ

ɲɲ rd ld lɲ jɲ lnd]. While [nɲ lɲ jɲ] are accidental gaps on Dixon's analysis, the other members of this set expand the phonotactic system in virtue of their initial non-apical nasals and clustering apical stops.

The constraints learned by our model capture the major generalizations on Wargamay consonant clusters and adjudicate the marginal cases in way that is sensitive to frequency of occurrence.[27]

(37) *Constraints on intervocalic consonant clusters*

Constraint	*Weight*	*Comment*
*[–son]C	7.27	The first member of a cluster must be a sonorant.
*C[+appr,–syl]	6.78	The second member of a cluster must not be [r l ɻ j].
*[+dors][+cor]	5.38	No dorsal-coronal clusters
*[+lab][–son,+cor]	5.35	No labial-[d ɟ] clusters (allows [mg mɲ])
*[–ant][+ant]	4.83	*[ɲ ɻ j][d n r l] (allows [ɲɟ jɟ])
*[–appr,–ant][+lab]	4.83	*[ɲ ɟ][+labial] (allows [nb nm])
*[+dors][+lab]	4.77	No dorsal-labial clusters
*[+lab][+dors]	4.47	No labial-dorsal clusters
*[–lat][+son,+ant]	3.32	*[r ɻ][n r l]
*CC[–syl,+son]	2.92	*CC[+sonorant]
*[+dors][–syl,+son]	2.33	Allows [ŋg]
*[–appr][+son,+ant]	2.27	Allows [j] before [n]
*[+lab][+son,–syl]	2.17	Allows [mg]
*[–appr,–ant][+son,+dors]	2.13	*[ɲ ɟ] before [ŋ]
*[+back,–syl]C	1.93	No [w]-initial clusters
*[–cons,–syl][+ant]	1.88	No glides before apico-alveolars
*[+lab][+son,+lab]	1.20	*[mm] (allows [mb])

Of the approximately 2400 two- and three- consonant clusters that are logically possible given the Wargamay segment inventory, these constraints assign perfect scores to only 51. The clusters predicted to be perfect include all of the homorganic nasal-stop sequences (36a), all of the clusters of type (36b), and the clusters of type (36c) that contain homorganic nasal-stop sequences (i.e., [rmb rɲɟ rŋg lmb lɲɟ lŋg ɻmb ɻɲɟ ɻŋg jmb jɲɟ jŋg]). In addition, the constraints assign perfect scores to some clusters that are not found in the learning data, namely [ɻm rŋg lŋg jŋg ɻŋg lnb lnɟ lng]. Of these, [ɻm lnb lnɟ lng] are accidental gaps according to Dixon's analysis.[28] The remaining clusters, [rŋg lŋg ɻŋg jŋg], can be rationalized as projections from the

[27] Recall that the learner was presented with *inflected* forms that do not contain any root/affix annotations. Further below we also discuss the constraints learned from the smaller set of roots.

[28] The constraints allow non-homorganic nasal-stop clusters after [l], but not after [r ɻ j], because [ln] is about 10 times more frequent than any of [rn ɻn jn]. That is, the system projects such triconsonantal clusters in proportion to the frequency of their biconsonantal components (cf. Pierrehumbert 1994).

attested triconsonantal clusters (which begin with [r l ɻ j]) and the frequent occurrence of [ɲg] in the learning data (resulting from the combination of a root ending in [ɲ] and the invariant dative/allative suffix [-gu]).

The constraints penalize some consonant clusters that are found in the learning data. Many of these are clusters that occur only under inflection ([mg mɲ nɲ ɲɲ lɲ]). The others are clusters that occur in just one or two roots ([nm nŋ rɲ rw lw ɻw jnb jɲ]).

In summary, there is a good numerical and qualitative fit between the clusters predicted by the learned grammar and Dixon's analysis. To the extent that the two differ, this can be attributed either to the fact that the set of clusters found in conjugated forms is larger than the set found in roots, or to a greater sensitive to frequency on the part of our model. The present analysis captures one major generalization that was not noted by Dixon, namely that non-homorganic sequences are marginal post-consonantally (especially after non-lateral sonorants).

As a further check of our model's ability to account for the consonant cluster inventory of Wargamay, we allowed it to learn constraints from the uninflected vocabulary items. The result, as expected, was a much tighter fit to this simpler portion of the language. Of the unattested clusters, only [ɻm rmb] were assigned perfect scores. And of the attested clusters, only rarely-occurring [rɲ rw lw ɻw nm nŋ jnb] were penalized.

8.3.4 Consonant-vowel combinations

In comparison to consonant cluster phonotactics, the regularities governing consonant-vowel combinations in Wargamay are understudied. However, Dixon (1981) does note one restriction on VC sequences: [ij] occurs pre-vocalically, but not before a consonant or at the end of the word. Further evidence for this phonotactic comes from the phonological rule of Yotic Deletion (Dixon 1981: 23), which eliminates [j] in the environment [i] { C,# } . Our model learns three constraints to cover this part of the system:

(38) *Constraints for yotic deletion*

Constraint	*Weight*	*Comment*
*[–back,+syl][–cons][^–long,+back]	3.87	*[i][jw]C, *[i][jw][i]
*[+high,+syl][^+son,+cor]#	3.64	*[iu]C#, where C ∉ [ŋɲrlɻ]
*[–back,+syl][^+son,+cor]#	1.79	*[i]C#, where C ∉ [ŋɲrlɻ]

Notice how the learner, in its rigorous pursuit of general constraints (§4.2.2), goes beyond Dixon's narrow description of the [ij] phonotactic. In the learning data, there are no instances of [iw]C (recall that [w] cannot appear in the first position of a consonant cluster) or [i][jw][i]; the first constraint folds these gaps together with the ban on [ij]C. Similarly, there are only nine roots in the vocabulary that exemplify [uj]#; the second constraint therefore expresses a gradient prohibition on both [ij]# and [uj]#, while the third constraint ensures that unattested [ij]# receives a greater penalty. (The complement class [^+son,+cor] appears in these constraints because it is the largest class that contains all of the legal word-final consonants except [j].)

8.4 Metrical phonotactics

Wargamay stress respects a distinction between heavy and light syllables, where a heavy syllable is defined as one containing a long vowel, irrespective of whether it is closed. Words containing all light syllables exhibit a right-to-left trochaic system:

(39) *Stress pattern of light-syllable words in Wargamay*[29]

ˈσ ŏ	[ˈbada]	‘dog’
ŏ ˈσ ŏ	[gaˈgara]	‘dilly bag’
ˈσ ŏ ˌσ ŏ	[ˈgiɟaˌwulu]	‘freshwater jewfish’
ŏ ˈσ ŏ ˌσ ŏ	[baˈɟinɟiˌlaŋgu]	‘spangled drongo-ERG/INSTR’
ˈσ ŏ ˌσ ŏ ˌσ ŏ	[ˈjajimˌbaliˌlagu]	‘play about-INTR.PURP’

As is evident from these examples, primary stress falls on the leftmost stressed syllable, following End Rule Left.

Heavy syllables (i.e., syllables with a long vowel) are limited to word-initial position in Wargamay, and all heavy syllables bear primary stress. Even-syllable words containing heavies exhibit the same stress pattern as all-light words. But three-syllable words of this type contain a lapse (sequence of unstressed syllables), because polysyllables never have final stress (Dixon 1981: 20).

(40) *Stress pattern of heavy-syllable words in Wargamay*

ˈσ̄ ŏ	[ˈmuːba]	‘stone fish’
ˈσ̄ ŏ ŏ	[ˈgiːbaɻa]	‘fig tree’ (*[ˈgiːbaˌɻa], [ˈgiːˌbaɻa])
ˈσ̄ ŏ ŏ ˌσ ŏ	[ˈguːŋaˌɻaɲiɲ]	‘rubbish-ABL’

There are no six-syllable words that begin with a heavy syllable in the training data (let alone Dixon’s vocabulary), and only one such word with five syllables. The stress pattern of this last form, [baːlbalilagu] ‘roll-INTR.PURP’, is uncertain: in particular, Dixon’s description does not make clear whether there is a lapse after the heavy syllable ([ˈbaːlbaliˌlagu]) or at the end of the word ([ˈbaːlbaˌlilagu]). We selected the former, based on pattern congruity, but will not consider such forms any further in light of our uncertainty about the facts.

To summarize, Wargamay has an essentially right-to-left trochaic stress pattern, with primary stress on the leftmost stressed syllable. Heavy syllables are limited to initial position,

[29] Dixon (1981:20) explicitly describes the stress pattern of words up to five syllables. We make the straightforward assumption that six-syllable words follow the same alternating pattern.

and three-syllable words that begin with a heavy have a final lapse. The learned grammar contains the following constraints on stress and length:

(41) *Constraints on stress and length*

Constraint	*Projection*	*Weight*	*Comment*
*# #	Main	1.75	Culminativity (top grid level)
*[+str][+main]	Stress	18.83	End Rule Left
*# #	Stress	0.94	Culminativity (middle grid level)
*[+str][+str]	Weight	16.01	*CLASH
*[–long][–str][–str]	Weight	11.97	No lapse after light
*[ ][+long]	Weight	9.19	No non-initial heavy
*[+main,–long]	Weight	7.36	No non-primary heavy
*#[–main][–main]	Weight	6.30	Initial window for main stress
*[–long,+str]#	Weight	3.40	No final stress on light
*[+str]#	Weight	2.71	No final stress
*##	Weight	0.94	Culminativity (lowest grid level)
*#[–main]#	Weight	0.94	Culminativity (lowest grid level)

We tested this set against all possible strings of up to length six of the set of possible syllables [ˈga ˈgaː ˌga ˌgaː ga gaː]. ([g] and [a] were chosen to prevent any distracting segmental violations.) As our test showed, the constraints of (41) assign penalties (of at least 6.23) to all incorrectly-stressed words and perfect scores to all correctly-stressed words except one, [ˈgaː], which received 2.78. The reason for this penalty was that there are only 15 monosyllabic words (all heavy) in the learning data (<1% of the total). No evidence is available concerning whether Wargamay speakers felt the few monosyllables of their language to be moderately aberrant.

We experimented with learning Wargamay without projections, and discovered that without a Weight projection the stress pattern was inaccessible, owing to the non-locality of the vowels (which were assumed to be the stress-bearing units). This is essentially the same reason why Shona vowel harmony was unlearnable without projections (§6.2). The system *could* learn Wargamay stress without the higher grid projections (Main and Stress, needed for unbounded stress; §7.2), though the resulting grammar was rather more complicated.

8.5 Additional phonotactics

The 43 constraints discussed above account for all of the phonotactics of Wargamay discussed by Dixon (1981). The learner also selected 49 additional constraints that have no direct analogue in Dixon's analysis.[30] A complete list of these constraints appears in Appendix B. Here we discuss a few types that have relatively straightforward interpretations.

The learned grammar contains two minimal-word constraints, one on the Vowel projection (*##, weight = 0.94) and one on the segmental projection (*#[ ]#, weight = 0.94). The small

[30] Hence, there were eight constraints, not reported here, that were discovered but ultimately weighted at zero; see fn. 7.

weights of these constraints make sense given the culminativity and other stress constraints that also penalize short words.

The learner selected a number of constraints that are violated by specific CV or VC sequences. Many of these refer to [i] or [iː]; they appear to be motivated by the relative rarity of these vowels, which are about half as frequent in the learning data as [u a] and [uː aː], respectively. For example, the constraint *[+cons,+son][+long,–back] (weight = 2.70) is violated by [mŋɲŋrlɻ][iː], and indeed all such sequences in the learning data are due to one root ([ŋiːɻa] 'tie up'). Similarly, *[+long,–back][+ant,–lat] (weight = 1.38) is violated by [iːr], and there are no such sequences in the vocabulary or learning data.

On the one hand, constraints such as these highlight the comprehensiveness of machine-learned phonotactic grammars, which will often contain constraints that would not be considered in traditional phonotactic analysis. On the other hand, their high degree of specificity may point to a weakness in our method of constraint selection, which does not directly take into account the frequency of individual segments. It may be that, given the baseline frequencies of [i], [iː], [r], and the other segments mentioned by the constraints, the number of expected sequences is roughly equal to the number that is observed. This suggests a question for future research: will such constraints still be selected if the learner is initially supplied with a constraint against every possible segment (and thus with the *a priori* ability to use individual-segment frequencies in the computation of sequence expectations)?

Finally, the fact that a small set of roots and affixes was used to construct the learning corpus results in a blurring of the line between phonotactic regularities and lexical entries. For example, there are two roots in Wargamay that contain a long vowel followed by [r] (i.e., [ŋaːra] 'hear, listen' and [ɟuːra] 'rub, wipe'). Because both of these roots have a short low vowel following [r], and the initial heavy syllable prevents that vowel from ever being stressed, the model constructs the unlikely constraint *[+long][+ant,–lat][–long,–high,–str] (weight = 3.04). Though some constraints at this level of detail might be eliminated by taking segment-specific frequency into account, as suggested above, a more straightforward way to avoid them would be to direct future work toward languages whose root corpora are substantially larger (and thus less prone to accidental gaps).

8.6 Summary

The present investigation of Wargamay has demonstrated the ability of our model to account for an entire phonotactic system. It has also sharpened Dixon's (1981) description of the language's segmental phonotactics, revealing gradient patterns in the word-final consonant inventory and a previously unnoticed restriction on non-homorganic nasal-stop clusters—and demonstrated the ability of metrical projections to account for a weight-sensitive stress pattern. We have suggested that limitations of the analysis are due either to the rather small vocabulary available for the language, or possibly to the fact that our constraint-selection method is not sensitive to segment-specific frequencies.

9. General discussion

In sum, we claim to have developed a system that can learn a non-trivial portion of the phonotactics of natural languages, given only a modest amount of information in the form of a segment inventory, a feature system, and a projection set. In so doing, we have developed arguments that phonological representations must include apparatus similar to the vowel tier (§6) and the metrical grid (§7, §8). In this final section we discuss questions that arise from our study and outline directions for future work.

9.1 Should the constraints be ranked?

We have noticed in various places that our model could create more general grammars if the constraints could somehow be ranked as they are in Optimality Theory. For instance, in Shona, our model learned three separate constraints (repeated below) that forbid mid vowels that fail to follow a harmony trigger.

(42)

	Constraint	*Projection*	*Comment*
a.	**a* [–high, –low]	Vowel	No mid vowels after *a*
b.	*[+high][–high, –low]	Vowel	No mid vowels after *i, u*
c.	*[–low, –back] *o*	Vowel	No *o* after *i, e*

In Optimality Theory, the usual analysis would be to unify these into a single constraint, perhaps as in (43).

(43) UNLICENSED MID: *[ ][–high,–low] (Vowel projection)

This forbids any mid vowel in a non-initial syllable, and would be outranked by the constraints requiring harmony, *[–high, –low] *i*, **o u*, and **e i*. This captures the intuition "noninitial mid vowels are bad unless they are the result of harmony".

Ranking also frequently seems plausible as applied to stress systems. A simple example arises in penultimate stress languages that tolerate (and assign stress to) monosyllables. Our learner responds to this pattern with the constraints in (44).

(44) *Constraints for penultimate stress in a language with stressed monosyllables*

	Constraint	*Projection*	*Comment*
a.	*# #	Main	Every word bears a main stress.
b.	*[ ][+stress] #	Segmental	Stress may not be final in polysyllables.

Under a ranking regime, we could simplify *[ ][+stress]# to just *[+stress]# (NONFINALITY; Prince and Smolensky 1993:42), by ranking it below *# #. This ranking expresses the intuition that final stress is avoided unless it would produce a stressless word. Similar cases can be found throughout Appendix A.

We concede a lesser elegance in such cases, but not necessarily the scientific ground, for the following reasons.

First, as Chomsky and Halle (1968:331) emphasize, we can only evaluate an acquisition model ((3)) by "confronting it with empirical evidence relating to the grammar that actually underlies the speakers performance"; they go on to say "we stress this fact because the problem has so often been misconstrued as one of 'taste' or 'elegance"'. We agree that the question is an empirical one. Phonotactic grammars that are insufficiently general typically leave gaps: illegal forms that fall between the constraints and are thus misclassified as legal. However, this has not been a problem for the grammars learned by our system. As we have shown, testing through exhaustive search, these grammars effectively separate well-formed structures from those that are ill-formed, with overlap limited to attested forms that are highly underrepresented. The reason is that the model is designed to defend actively against gaps. The process of sample creation (§3.3) constantly explores the space of phonotactic possibilities, looking for illegal forms that can be ruled out with new constraints.

The other reason to favor general grammars is that only such grammars can account for how humans extend their knowledge to new forms. For instance, a grammar of English that simply listed the existing syllable onsets would fail to generalize to unattested onsets in the way observed by Scholes (1966). Because our model seeks general constraints (§4.2.2) based on natural classes, it captures the distinctions among the unattested clusters tested by Scholes rather well. Whether the model would perform even more accurately if it made use of constraint ranking is a matter for future work to determine.

Moreover, there are considerable advantages arising from using a non-ranked maximum entropy approach instead of Optimality Theory, advantages that become clear when we examine existing work. First, unlike the OT phonotactic learning algorithms of Hayes (2004) and Prince and Tesar (2004), our model has the ability to learn gradient well-formedness (§2.3) on the basis of underrepresentation in the learning data. The models just cited work with heuristics designed to rank Markedness constraints above Faithfulness constraints wherever possible. These heuristics apply on an all-or-nothing basis, creating strictly ranked phonotactic grammars. Such grammars are "brittle," by which we mean two things: they cannot rate forms gradiently (going against experimental observation; §2.3), and they cannot learn a constraint if it has even one counterexample. Thus, for example, one single token of [pw] (e.g. *Puerto Rico*), would suffice to sink the English constraint against labial + [w] clusters. In contrast, a maximum entropy model responds flexibly and sensitively to the range of frequencies encountered in the learning data.

The OT model of Pater and Coetzee (2006) does have the capacity to treat gradience, reacting to imperfect phonotactic generalizations by creating lexically-specific Faithfulness constraints. However, the statistic this model employs is just O (Observed), not O/E (Observed/Expected)—essentially, it ranks Markedness constraints by sorting them in increasing order of O. This is problematic, because constraints with identical O values but sharply different E values differ greatly in their effects. For example, Clements and Keyser (1983:48) propose a constraint whose sole purpose is to ban the onset [stw]. Since English lacks [stw], this constraint has an O value of zero. The E value for this constraint surely is low, since [stw] contains [tw], which is rare ((12)) and indeed is already penalized by a constraint ((14.24)) discovered by our learner. In contrast, the onset [skt] violates a very general constraint on sonority sequencing; in our simulation, this is the highly-weighted (14.3). This constraint also has an O of zero, but because it is not reducible to simpler constraints, it has a much larger E value. While we lack

experimental data, we think it very likely that [skt] would be rated as worse than [stw]. Such cases suggest that O alone will not suffice to model native intuition; E is needed, too.

The stochastic version of OT (Boersma 1997) can capture gradient effects in phonological alternation (Boersma and Hayes 2001, Hayes and Londe, in press), and could in principle be applied to phonotactics. However, it faces a major conceptual problem, pointed out by Pater and Coetzee (2006): for marginal forms, stochastic ranking of the relevant Markedness and Faithfulness constraints predicts not marginality, but free variation with some less-marked alternative (for *Puerto Rico* these might be [ˈportoʊ], [pu.ˈwɛrtoʊ], etc.). But often, speakers use forms invariantly even when they judge them to be deviant.

In conclusion, we judge that it would be premature to rule out ranking as a basis for phonotactics, particularly since we do appreciate ourselves the elegance of descriptions attainable with ranking. We thus would find it of interest if there is a feasible mathematical scheme that could increase the generality of constraints through ranking, while retaining the quantitative robustness and accuracy of the maximum entropy system.

9.2 Hidden structure

In recent work Tesar and Smolensky (1998, 2000) and Tesar (2004) have addressed the role of "hidden structure" in phonological learning. By this they mean structure that is not detectible in the phonetic signal, but which arguably is phonologically present and provides order and systematicity to the phonological pattern. An example of hidden structure is syllable weight (e.g. Hayes 1995:§3.9.2): certain properties of a syllable are used to classify syllables into light and heavy categories, which then can be used to make sense of other patterns, particularly stress.

Hidden structure is often language-specific; for example, different languages impose different criteria for what counts as a heavy syllable. This creates a "chicken-or-egg" problem: we need to know the language-specific criterion of syllable weight to be able to detect the stress pattern, but it is often the stress pattern itself that gives the main evidence for the syllable weight criterion. Tesar and Smolensky offer intriguing methods, based on expectation maximization and inconsistency detection, to discover both the hidden structure and the generalizations based on it.

While our present model incorporates no clear cases of hidden structure, we believe it could be scaled up to learn it. The crucial idea is that when the right hidden structure is selected, this can be detected by the methods of maximum entropy. Specifically, correct hidden structure leads to a tighter phonotactic characterization, which increases the probability of the learning data, a measurable quantity under the maximum entropy approach. In future work we hope to address the problem in these terms.

9.3 Relating phonotactics to alternation

Phonological alternation occurs when morphemes take on different forms in different environments. It is related to phonotactics because alternations frequently are seen to enforce the phonotactics dynamically. For instance, the English plural morpheme /-z/ is altered to [-s] following voiceless obstruents, as in *cups* [kʌps], in order to avoid a violation of the phonotactic constraint that bans voicing agreement in final obstruent clusters. This is the essence of the

"conspiracy problem" (Kisseberth 1970), which has been the focus of a great deal of phonological theorizing, notably in Optimality Theory. A central goal of OT is to reduce the description of alternations to the same principles that govern phonotactics. In particular, the ranking of Markedness over Faithfulness constraints results in both static restrictions on surface forms and alternations that respect those restrictions.

Despite the many successes achieved in OT, we are not convinced that its method for linking alternations to phonotactics is correct. For instance, it is difficult to extend it to cases where alternation does not address phonotactic problems. Thus, for instance, in Yidiɲ phonology, [u] is chosen (productively) as the epenthetic vowel following a nasal consonant, yet there is no evident connection between nasality and [u] in Yidiɲ phonotactics (Hayes 1999b). English vowel length alternations *(SPE*) are phonotactically motivated insofar as they optimize foot structure (see Prince 1990, Hayes 1995), but the accompanying quality alternations ([iː] ~ [ɛ], [eɪ] ~ [æ], [aɪ] ~ [ɪ], [oʊ] ~ [ɑ]) have no evident phonotactic basis. The candidate **prof*[ɛ]*nity* for *prof*[eɪ]*n* + *ity* would be more Faithful than *prof*[æ]*nit*y, and phonotactically just as good. The point is that if humans posses some mechanism for learning alternations that are phonotactically unmotivated, it is difficult to understand why that mechanism would not also be used in learning the phonotactically motivated alternations as well.

We suggest that the proper link between alternations and phonotactics is at the level of language learning: knowing the phonotactics makes it easier for the language learner to discover alternations, because she expects *a priori* that alternation should occur for this purpose. Thus, for example, an English-learning child who already knew the principle of voicing agreement in final obstruent clusters would be in a good position to understand and analyze the voicing alternation in the plural suffix (Albright and Hayes 2002, Hayes 2004, Prince and Tesar 2004). That is, it would be immediately apparent that the simple concatenation [kʌps]+[z] is insufficient for the plural of *cup*, owing to its phonotactic violation; and it would remain only to find the change needed produce the correct output [kʌps].

There is experimental evidence compatible with this conception. Children evidently learn at least some of the phonotactics of their language very early (i.e., in infancy; see Hayes 2004 for literature review)—so that whatever model of acquisition we develop should in any event include the capacity to learn phonotactics solely from distributional data. Moreover, the experimental findings of Pater and Tessier (2003) suggest that phonotactic knowledge does indeed assist learners in finding alternation patterns. A learning system for phonological alternations devised by Albright and Hayes (2002, 2003) already incorporates an elementary capacity to use phonotactic knowledge to assist learning, as does the OT-based system of Tesar and Prince (to appear).

In sum, as a tentative answer to the question of how the work described here should be related to the problem of phonological alternation, we suggest a long-term research program of *learning-theoretic phonology*. Specifically, we advocate an architecture for phonological theory that recapitulates the process by which it is learned, and learning systems that can steadily make use of what has been learned so far to assist at each successive stage of learning.

9.4 How is phonological typology to be explained?

While establishing the content of the acquisition module AM ((3) above) strikes us as the central theoretical challenge in phonology, there is a second question that also deserves attention: why are languages the way they are? More specifically, what is the basis for the systematic cross-linguistic patterning, especially involving markedness, that emerges from typological study? Certainly an inductive-baseline learner will provide no explanations for these patterns; those typologically unnatural patterns that can be characterized by general and accurate constraints will be just as learnable as the typologically natural ones.

One possible response to this question would be to say that, as our inductive baseline strategy is pursued further, it will turn out that the only effective learning strategy is one with an extremely rich UG—a UG that incorporates the entire constraint set for phonology (Prince and Smolensky 1993; Tesar and Smolensky 1998, 2000; McCarthy 2002). If so, the problem of typology will likely be solved, and our efforts in developing methods to extract good constraints from a large search space will turn out to have been merely a foil, used to provide complete vindication for the innate-constraint approach.

However, there are other ways to enrich the inductive baseline model that are rather more conservative in their reliance on innate knowledge. For instance, language learners could make use of their own phonetic experience, accessing it to discover phonetically natural constraints grounded in articulation and perception (Hayes 1999a; Steriade 1999, 2001a, b.; Gordon 2004; Hayes, Kirchner and Steriade 2004). Preference for such constraints would constitute a *learning bias* in favor of phonological systems that are easier to produce or perceive, or that suffer a lesser recognition burden from alternation. For experimental evidence in favor of learning biases, and a mechanism (based on maximum entropy) whereby they could be incorporated into a general learning scheme, see Wilson (to appear).

Further afield, we note that many scholars hold the view that not all typological patterns should be explained by UG; instead, the diachronic process of language transmission and mistransmission is responsible for much or all of typology. For representative statements of this idea, see Ohala (1981), Blevins (2004), and Myers (ms.). We think that more serious assessment of this position will become possible as formally implemented models of the process of language mistransmission (Kochetov 2002) become increasingly available.

9.5 Directions for future work

In expanding the approach taken here, we think an important line to follow will be to enrich the class of formal mechanisms it can access. In other words, while we have shown that vowel tiers and grids are important to phonotactic learning, we judge that our system is still far too close to its original inductive baseline, as there are phonological phenomena it clearly cannot learn unless further modified. We give two examples here.

First, we cannot claim that our system of projections has fully solved the problem of learning nonlocal phonotactic dependencies. Most notably, it cannot account for consonant-to-consonant dependencies of the kind studied in McCarthy (1979, 1988), MacEachern (1999), Frisch, Pierrehumbert, and Broe (2004), and Rose and Walker (2004). Simply adding a

consonant projection is unlikely to suffice for these cases, because of two special factors. Consonant-to-consonant phonotactics relies heavily on similarity (those consonants that are already most similar are the ones whose distribution is phonotactically regulated). Further, there are also gradient distance effects: consonants that are separated at a short distance are regulated more closely than those at greater distances. Neither of these effects could be modeled merely by introducing a consonant projection. We anticipate that the right approach would be to incorporate a similarity metric into the theory (perhaps that proposed by Frisch et al.) and use it to scan the nearby segments.

We also lack a theory to learn the phonotactics of neutral vowels; i.e. cases where particular vowels (not just consonants) are skipped over in vowel harmony. We are encouraged here by findings (Gordon 1999, Benus and Gafos, to appear) that in their allophonic forms, neutral vowels can be weakly harmonic, taking on slightly different phonetic forms depending on the neighboring harmonic vowels. The incorporation of such phonetic detail into the representations would "localize" the phonotactics on the vowel projection, perhaps sufficing to make neutral-vowel phonotactics learnable.

The above two lacunae in our approach are surely not the only ones: only extended study and modeling of many languages can show what is needed in a phonotactic learning model.

Appendix A: Training sets and constraints for the stress typology of Gordon (2002)

Constraints are listed in discovery order. Abbreviations: (M) = Main tier, (S) = Stress tier; otherwise Default tier; [s] = [stress], [m] = [main], 1 = [+main], 2 = [+stress, –main], 0 = [–stress].

Language	*Stress pattern*	*Constraints Learned*
Araucanian	1, 01, 010, 0102, 01020, 010202, 0102020, 01020202	*# # (M) 5.4, *[+s]1 (S) 6.3, *[+s][+s] 6.0, *00 6.3, *#[ ][-m] 4.9, *#[ ][ ][+s] 2.3
Atayal	1, 01, 001, 0001, 00001, 000001, 0000001, 00000001	*# # (M) 6.8, *[+s][+s] (S) 2.1, *[+s][ ] 7.6
Biangai	1, 10, 210, 2010, 22010, 202010, 2202010, 20202010	*# # (M) 5.1, *1[+s] (S) 6.6, *#0 5.4, *00 6.0, *[ ][+s]# 3.1, *[ ][+s][+s] 6.3, *[-m][ ]# 3.7
Cavinena	1, 10, 010, 2010, 02010, 202010, 0202010, 20202010	*# # (M) 5.4, *1[+s] (S) 6.3, *[+s][+s] 6.5, *00 6.3, *[ ][+s]# 3.1, *[-m][ ]# 3.7
Cayuvava	1, 10, 100, 0100, 00100, 200100, 0200100, 00200100	*# # (M) 5.8, *1[+s] (S) 5.4, *[+s][+s] 2.5, *[+s][+s][+s] (S) .8, *[ ][+s]# 4.6, *[+s][ ][+s] 2.3, *000 4.5, *[ ][+s][ ]# 3.0, *[-m][ ][ ]# 3.5, *[+s][ ][ ][-m] 4.2
Central Alaskan Yupik	1, 01, 021, 0201, 02021, 020201, 0202021, 02020201	*# # (M) 4.5, *1[+s] (S) 7.0, *[-m]# 5.6, *#2 6.1, *[+s]2 3.4, *00 5.4, *0[ ][+s][ ] 3.6
Chitimacha	1, 10, 100, 1000, 10000, 100000, 1000000, 10000000	*# # (M) 4.5, *[+s][+s] (S) 7.9, *#[-m] 7.3
Creek	1, 01, 010, 0201, 02010, 020201, 0202010, 02020201	*# # (M) 6.0, *1[+s] (S) 6.7, *[+s][+s] 4.2, *#2 2.1, *00 3.5, *#[+s][ ] 4.9, *0[ ][ ]0 3.7, *0[-m][ ]# 2.6
Estonian (data from Hint 1973)	1, 10, 100, 1020, 10200, 10020, 102020, 100200, 1020200, 1020020, 1002020, 10202020, 10200200, 10020200, 10020020	*# # (M) 4.5, *[+s]1 (S) 3.2, *#[-m] 3.6, *[+s][+s] 6.5, *2# 5.8, *000 5.6, *[ ]1 4.4
Garawa	1, 10, 100, 1020, 10020, 102020, 1002020, 10202020	*# # (M) 4.7, *[+s]1 (S) 3.1, *#[-m] 3.3, *[+s][+s] 6.3, *2# 5.9, *[-m]00 6.3, *[ ]1 4.7
Georgian	1, 10, 100, 0100, 20100, 200100, 2000100, 20000100	*# # (M) 5.6, *1[+s] (S) 3.0, *[+s][+s] 4.5, *[+s][+s][+s] (S) 2.9, *[ ]2 3.1, *[ ][+s]# 4.6, *#0[-m] 5.6, *[ ][+s][ ]# 2.5, *[-m][ ][ ]# 4.3, *1[ ][ ][ ] 4.0
Gosiute Shoshone	1, 12, 102, 1022, 10202, 102022, 1020202, 10202022	*# # (M) 4.5, *[+s]1 (S) 6.8, *#[-m] 6.1, *0# 5.6, *00 6.3, *[+s][+s][ ] 6.5
Hopi	1, 10, 010, 0100, 01000, 010000, 0100000, 01000000	*# # (M) 5.4, *[+s][+s] (S) 7.6, *#[-m][-m] 2.0, *#[ ][-m][ ] 6.2, *[ ][+s]# 4.4
Indonesian	1, 10, 010, 2010, 20010, 202010, 2002010, 20202010	*# # (M) 5.4, *1[+s] (S) 3.4, *[+s][+s] 6.1, *[ ][+s]# 3.1, *[-m][ ]# 3.7, *#[ ]2 1.3, *#0[-m] 5.6, *000 1.5, *[ ][ ]00 5.2, *1[ ][ ] 3.7
Ioway-Oto	1, 01, 010, 0100, 01002, 010020, 0100200, 01002002	*# # (M) 5.4, *[+s]1 (S) 5.8, *[+s][+s] 3.4, *#[ ][-m] 6.1, *[+s][ ][+s] 2.5, *000 2.7, *[+s][ ][ ]0 3.1, *0[ ][ ][+s] 3.4, *#[ ][ ][+s] 2.6
Lakota	1, 01, 010, 0100, 01000, 010000, 0100000, 01000000	*# # (M) 5.4, *[+s][+s] (S) 7.9, *#[ ][-m] 7.3
Lower Sorbian	1, 10, 100, 1020, 10020, 100020, 1000020, 10000020	*# # (M) 4.7, *[+s]1 (S) 2.5, *#[-m] 3.4, *[+s][+s][+s] (S) 3.1, *[+s][+s] 4.5, *2# 4.5, *2[ ][ ] 3.5, *[-m]0[ ]# 5.8, *[ ]1 4.8

Macedonian	1, 10, 100, 0100, 00100, 000100, 0000100, 00000100	*# # (M) 5.8, *[+s][+s] (S) 4.5, *[ ][+s]# 4.6, *[ ][+s][ ]# 3.3, *[+s][ ][ ][ ] 4.0, *[-m][ ][ ]# 4.0
Malakmalak	1, 10, 010, 1020, 01020, 102020, 0102020, 10202020	*# # (M) 6.3, *[+s]1 (S) 6.9, *[+s][+s] 5.5, *00 4.9, *2# 2.1, *[ ][+s]# 4.9, *0[ ][ ]0 2.8
Maranungku	1, 10, 102, 1020, 10202, 102020, 1020202, 10202020	*# # (M) 4.8, *[+s]1 (S) 3.0, *#[-m] 3.3, *[+s][+s] 6.8, *00 6.5, *[ ]1 4.3
Nahuatl	1, 10, 010, 0010, 00010, 000010, 0000010, 00000010	*# # (M) 6.8, *[+s][+s] (S) 2.6, *[ ][+s]# 6.3, *[+s][ ][ ] 7.2
Pacific Yupik	1, 01, 010, 0102, 01002, 010020, 0100202, 01002002	*# # (M) 5.4, *[+s]1 (S) 3.4, *[+s][+s] 6.2, *#[ ][-m] 6.1, *00# 5.0, *[+s][+s][+s][+s] (S) .8, *000 5.4, *[+s][ ][+s][ ] 5.8, *[ ][ ]1 3.3
Palestinian Arabic	1, 10, 201, 2010, 20201, 202010, 2020201, 20202010	*# # (M) 5.2, *1[+s] (S) 6.6, *[+s][+s] 5.2, *#0 6.1, *00 4.3, *[+s][ ][ ][+s] 3.1, *[-m][-m]# 2.6
Pintupi	1, 10, 100, 1020, 10200, 102020, 1020200, 10202020	*# # (M) 4.8, *[+s]1 (S) 3.0, *#[-m] 3.3, *[+s][+s] 4.8, *2# 5.3, *0[ ][+s] 1.8, *00[ ] 2.9, *[ ]1 4.3, *[+s][ ]0[ ] 3.4
Piro	1, 10, 010, 2010, 20010, 202010, 2020010, 20202010	*# # (M) 5.4, *1[+s] (S) 3.4, *[+s][+s] 5.9, *[ ][+s]# 2.9, *[-m][ ]# 3.9, *#[ ]2 2.1, *#0[-m] 4.8, *00[-m] 5.6, *1[ ][ ] 4.1
Quebec French	1, 21, 201, 2001, 20001, 200001, 2000001, 20000001	*# # (M) 4.6, *1[+s] (S) 2.3, *[-m]# 3.5, *[+s][+s][+s] (S) 3.7, *#0 6.1, *1[ ] 4.8, *[ ]2 4.0
Sanuma	1, 10, 010, 2010, 20010, 200010, 2000010, 20000010	*# # (M) 5.3, *1[+s] (S) 6.3, *[+s][+s] 4.4, *[+s][+s][+s] (S) 3.2, *[ ]2 3.4, *[ ][+s]# .9, *[-m][ ]# 7.1, *#0[-m] 5.8
Southern Paiute	1, 10, 010, 0120, 01020, 010220, 0102020, 01020220	*# # (M) 5.4, *[+s]1 (S) 6.3, *00 6.1, *2# 2.3, *#[+s][+s] .5, *[+s][+s][+s] 1.1, *#[ ][-m][ ] 3.6, *[ ][+s]# 4.9, *[+s][+s][ ][ ] 5.7, *#[+s][ ][ ] 3.0
Tauya	1, 21, 201, 2201, 20201, 220201, 2020201, 22020201	*# # (M) 4.8, *1[+s] (S) 3.2, *[-m]# 3.2, *#0 5.6, *00 6.3, *[ ][+s][+s] 6.5, *1[ ] 4.4
Udihe	1, 01, 201, 2001, 20001, 200001, 2000001, 20000001	*# # (M) 4.6, *1[+s] (S) 2.5, *[+s][+s] 4.4, *[+s][+s][+s] (S) 3.4, *[-m]# 3.5, *[ ]2 3.7, *#0[-m] 6.0, *1[ ] 4.9
Urubu Kapor	1, 01, 201, 0201, 20201, 020201, 2020201, 02020201	*# # (M) 4.8, *1[+s] (S) 3.3, *[+s][+s] 4.9, *[-m]# 2.9, *00 5.2, *[+s][ ][ ][+s] 2.4, *1[ ] 4.0, *0[ ][ ]# 1.8
Walmatjari (data from Hudson 1978)	1, 10, 100, 1020, 10200, 10020, 100200, 100020, 1000200, 1000020, 10000200, 10000020	*# # (M) 4.1, *[+s]1 (S) 6.0, *#[-m] 6.9, *[+s][+s][+s] (S) 6.0, *[+s][+s] 4.7, *2# 4.1, *00[ ]# 6.1
Winnebago	1, 01, 001, 0010, 00102, 001002, 0010020, 00100202	*# # (M) 5.4, *[+s]1 (S) 5.8, *[+s][+s] 5.8, *#[+s][ ] 2.7, *[-m]00 2.6, *[+s][+s][+s][+s] (S) .1, *00# 4.5, *#[ ][ ][-m] 5.9, *[+s][ ][+s][ ] 2.3, *[+s][ ][ ]0 3.6, *#[ ][-m]# 2.4

Appendix B: Constraints for Wargamay not discussed in the text

All constraints listed were discovered on the Default projection.

Constraint	*Weight*
*[+high,–main,+str][–lat]	6.04
*[–main,+str][–son,+lab]	5.39
*[+son,+dors][–main,+str]	5.35
*[+ant,–lat][+high,–main,+str][–cons]	4.67
*[+high,+back,–main,+str][+son,+lab]	4.61
*[+son,+lab][–back,–main,+str]	4.51
*[–main][–son,+ant][+high,–main]	4.49
*[+high,+bk,–main][–son][–back,–main]	4.49
*V[+son,+dors][–back]	4.49
*[–main,+str][–son,+ant]	4.46
*[+son,–app][–long,+back,–str] #	4.44
*[–syl][+son,+cor][+back]	4.42
*[–app][–syl][–long]	4.37
*[–back][+high,–main,+str]	4.33
*[–main,+str][–cons][+back]	4.32
*[+app,–syl][+ant][+str]	4.22
*[+high,+back,–main,+str][–son,–ant]	3.94
*[–back,+syl][–cons][–long,+back]	3.93
*[+high,+syl][+son,+cor]#	3.70
*[–app][+son][+back,–main]	3.66
*[+high,+back][–app] #	3.59
*[–back,–main][+back]	3.52
*# [+ant][–long,+back]	3.45
*V[+back][–long,+back]	3.04
*[+long][+ant,–lat][–long,–high,–str]	2.88
*[+cons,+son][+long,–back]	2.86
*[–app][+son][+high,+str]	2.83
*[–str][+ant][+back,–str]	2.76
*[–high][–cons][+ant]	2.71
*[–main][+back][–long,+back,+str]	2.68
*# [+cons,+app][–long,+back]	2.65
*[+back][+long][+app]	2.53
*[–cons][+long,+high][–app,+cor]	2.51
*[+long][+lat][+son,–syl]	2.49
*[+son,+cor][+long]	2.42
*[–son,–ant][+long][–cons]	2.39
*[–back][+high,+syl][–cons]	2.17
*[–back][+long,–high]	1.99
*[–back,+syl][+son,+cor]#	1.82
*[+long,–back][+son,+lab]	1.80
*[–back][+long][+cons,+son]	1.74
*[+long,+high][–cons][–long,–back,–str]	1.65
*[–syl][+son,+cor][+high,+back]	1.59
*[+back]V[+back]	1.53
*[+long,–back][–son,+ant]	1.53

*[–syl][+son,+cor][+high,+main]	1.43
*[+long,–back][+ant,–lat]	1.40
*[+str][+lat] #	1.34
*[+long][+back]	1.31

References

Albright, Adam. 2002. The identification of bases in morphological paradigms. Doctoral Dissertation, University of California, Los Angeles.

Albright, Adam. 2006. Gradient phonotactic effects: lexical? grammatical? both? neither? Talk presented at the 80th Annual Meeting of the Linguistic Society of America.

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the 2002 Workshop on Morphological Learning, Association of Computational Linguistic*s, ed. Mike Maxwell. Philadelphia: Association for Computational Linguistics.

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90:119–161.

Albro, Daniel M. 1998. Evaluation, Implementation, and Extension of Primitive Optimality Theory. Master's thesis, University of California, Los Angeles.

Albro, Daniel M. 2005. Computational Optimality Theory and the Phonological System of Malagasy. Doctoral Dissertation, University of California, Los Angeles.

Allauzen, Cyril, Mehryar Mohri, and Brian Roark. 2005. The design principles and algorithms of a weighted grammar library. *International Journal of Foundations of Computer Science* 16:403– 421.

Archangeli, Diana. 1984. *Underspecification in Yawelmani Phonology and Morphology*. New York: Garland.

Archangeli, Diana, and Douglas Pulleyblank. 1987. Maximal and minimal rules: effects of tier scansion. In *Proceedings of the North Eastern Linguistic Society 17*, ed. Joyce McDonough and Bernadette Plunkett, 16–35. Amherst: Graduate Linguistics Student Association, University of Massachusetts.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 44:568–591.

Beckman, Jill. 1997. Positional faithfulness, positional neutralisation and Shona vowel harmony. *Phonology* 1–46.

Benus, Stefan, and Adamantios Gafos. to appear. Articulatory characteristics of Hungarian "transparent" vowels. *Journal of Phonetics* .

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

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound pattern*s. Cambridge: Cambridge University Press.

Bloomfield, Leonard (1933) *Language*. New York: Henry Holt.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Institute of Phonetic Sciences, University of Amsterdam, Proceedings 2*1, 43–58.

Boersma, Paul. 2004. A Stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments. Rutgers Optimality Archive 648, http://roa.rutgers.edu.

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

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10:425–455.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Chomsky, Noam. 1963. Formal properties of grammars. In *Handbook of Mathematical Psychology*, ed. R. D. Luce, R. R. Bush, and E. Galanter, volume II, 323–418. New York: Wiley.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of Englis*h. Cambridge, MA: MIT Press.

Clements, George N. 1976. Neutral vowels in Hungarian vowel harmony: an autosegmental interpretation. In *Proceedings of the Seventh Annual Meeting of North Eastern Linguistic Societ*y, ed. Judy Kegl, David Nash, and Annie Zaenen, 49–64.

Clements, George N., and Elizabeth V. Hume. 1995. The internal organization of speech sounds. In *The Handbook of Phonological Theor*y, ed. John Goldsmith, 245–306. Oxford: Blackwell.

Clements, George N., and S. Jay Keyser. 1983. *CV Phonology: A Generative Theory of the Syllabl*e. Cambridge, MA: MIT Press.

Clements, George N., and Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. In *The Structure of Phonological Representations (part II*), ed. Harry van der Hulst and Norval Smith, 213–256. Dordrecht: Foris.

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonolog*y, 49–56. Somerset, NJ: Association for Computational Linguistics.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:380–393.

Dixon, Robert M. W. 1981. Wargamay. In *Handbook of Australian Languages, Volume II*, ed. Robert M. W. Dixon and Barry J. Blake, 1–144. Amsterdam: John Benjamins.

Dresher, B. Elan, and Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 20:421–451.

Eisner, Jason. 1997. Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic*s, 313–320.

Eisner, Jason. 2001. Expectational semirings: Flexible EM for finitestate transducers. In *Proceedings of the ESSLLI Workshop on Finite-State Methods in NLP (FSMNLP*), ed. G. van Noord.

Eisner, Jason. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic*s, 1–8.

Ellison, T. Mark. 1994. Phonological derivation in Optimality Theory. In *Proceedings of the Fifteenth International Conference on Computational Linguistic*s, 1007–1013.

Fortune, G. 1955. *An analytical grammar of Shona*. London: Longmans, Green & Co.

Frisch, Stefan A. 1996. Similarity and frequency in phonology. Doctoral Dissertation, Northwestern University. URL http://www.cas.usf.edu/ frisch/publications.html.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42:481–496.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179–228.

Frisch, Stefan A., and Bushra A. Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77:91–106.

Fudge, Erik C. 1969. Syllables. *Journal of Linguistics* 253–287.

Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological rule induction. *Computational Linguistics* 22:497–530.

Goldsmith, John. 1979. *Autosegmental Phonolog*y. New York: Garland.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theor*y, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

Gordon, Matthew. 1999. The "neutral" vowels of Finnish: How neutral are they? *Linguistica Uralica* 35:17–21.

Gordon, Matthew. 2002. A factorial typology of quantity insensitive stress. *Natural Language and Linguistic Theory* 20:491-552.

Gordon, Matthew. 2004. Syllable weight. In *Phonetically Based Phonology*, ed. Bruce Hayes, Robert Kirchner, and Donca Steriade, 277–312. Cambridge: Cambridge University Press.

Greenberg, Joseph H., and J. J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20:157–177.

Hale, John, and Paul Smolensky. 2006. Harmonic Grammars and harmonic parsers for formal languages. In *The Harmonic Mind: from Neural Computation to Optimality-Theoretic Grammar*, ed. Paul Smolensky and Géraldine Legendre, chapter 10, 393-416. Cambridge, MA: MIT Press.

Halle, Morris. 1959. *The Sound Pattern of Russian*. The Hague: Mouton.

Halle, Morris and George N. Clements. 1983. *Problem book in phonology*. Cambridge, MA: MIT Press.

Halle, Morris, and Jean-Roger Vergnaud. 1987. *An Essay on Stress*. Cambridge, MA: MIT Press.

Hammond, Michael. 1999. *The phonology of English: a prosodic Optimality-theoretic approach*. Oxford: Oxford University Press.

Hammond, Michael. 2004. Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies* 4:1-24.

Hannan, M. 1959. *Standard Shona dictionary*. New York: St Martin's Press.

Hannan, M. 1981. *Standard Shona dictionary, 2nd edition with Addendum*. Salisbury, Harare: The Literature Bureau.

Hay, Jennifer, Janet B. Pierrehumbert, and Mary Beckman. 2003. Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in Laboratory Phonology VI*, ed. John Local, Richard Ogden, and Rosalind Temple, 58–74. Cambridge: Cambridge University Press.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

Hayes, Bruce. 1999a. Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In *Functionalism and Formalism in Linguistics, Volume I: General Papers*, ed. Mike Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatley, 243–285. Amsterdam: John Benjamins.

Hayes, Bruce. 1999b. Phonological restructuring in Yidiŋ and its theoretical consequences. In *The Derivational Residue in Phonological Optimality Theory*, ed. Ben Hermans and Marc Oostendorp, 175–295. Amsterdam: John Benjamin.

Hayes, Bruce. 2000. Gradient well-formedness in Optimality Theory. In *Optimality Theory: Phonology, Syntax, and Acquisition*, ed. Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer, 88-120. Oxford: Oxford University Press.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In *Fixing Priorities: Constraints in Phonological Acquisition*, ed. René Kager, Joe Pater, and Wim Zonneveld, 158-203. Cambridge University Press.

Hayes, Bruce, Robert Kirchner, and Donca Steriade, eds. 2004. *Phonetically-based phonology*. Cambridge: Cambridge University Press.

Hayes, Bruce, and Zsuzsa Cziráky Londe. In press. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23.

Heinz, Jeffrey. to appear-a. Learning phonotactic patterns from surface forms. In *Proceedings of The 25th West Coast Conference on Formal Linguistics (WCCFL 25*).

Heinz, Jeffrey. to appear-b. Learning quantity-insensitive stress patterns via local inference. In *Proceedings of The Association for Computational Linguistics Special Interest Group in Phonology 6 (ACL-SIGPHON 06*).

Hint, Mati. 1973. *Eesti Keele Sonafonoloogia* I. Tallinn, Estonia: Eesti NSV Teaduste Akadeemia.

Hopcroft, John E., and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Hudson, Joyce. 1978. *The Core of Walmatjari Grammar*. Canberra: Australian Institute of Aboriginal Studies.

Jäger, Gerhard. 2004. Maximum entropy models and stochastic Optimality Theory. Ms., University of Potsdam.

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

Jaynes, Edwin T. 1983. *Papers on Probability, Statistics, and Statistical Physic*s. D. Reidel Publishing Company.

Jelinek, Frederick. 1999. *Statistical Methods for Speech Recognitio*n. Cambridge, MA: MIT Press.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognitio*n. Upper Saddle River, NJ: Prentice Hall.

Kager, René. 1995. The metrical theory of word stress. In *The Handbook of Phonological Theor*y, ed. John Goldsmith, 367–402. Oxford: Blackwell.

Keller, Frank. 2005. Linear Optimality Theory as a model of gradience in grammar. In *Gradience in Grammar: Generative Perspective*s, ed. Gisbert Fanselow, Caroline Fery, Ralph Vogel, and Matthias Schlesewsky. Oxford University Press.

Kiparsky, Paul. 1973. Phonological representations. In *Three Dimensions of Linguistic Theor*y, ed. Osamu Fujimura. Tokyo: TEC.

Kiparsky, Paul. 1982. Lexical Phonology and Morphology. In *Linguistics in the Morning Cal*m, ed. In-Seok Yang, 3–91. Seoul: Hanshin.

Kisseberth, Charles W. 1970. On the functional unity of phonological rules. *Linguistic Inquiry* 1:291–306.

Klein, Dan, and Christopher Manning. 2003. Maxent models, conditional estimation, and optimization, without the magic. Tutorial presented at NAACL-03 and ACL-03.

Kochetov, Alexei. 2002. *Production, perception, and emergence phonotactic patterns.* New York: Routledge.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: an application. In *COGSCI 199*0, 884–891.

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006. The Optimality Theory - Harmonic Grammar connection. In *The Harmonic Mind: from Neural Computation To Optimality-Theoretic Gramma*r, ed. Paul Smolensky and Géraldine Legendre, volume 2, 339–402. Cambridge, MA: MIT Press.

Liberman, Mark. 1975. The Intonational System of English. Doctoral Dissertation, Department of Linguistics and Philosophy, MIT. Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8:249–336.

Liberman, Mark and Alan S. Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8:249-336.

MacEachern, Margaret R. 1999. *Laryngeal cooccurrence restriction*s. New York: Garland. MacKay, David J. C. 2003. *Information theory, inference, and learning algorithm*s. Cambridge: Cambridge University Press.

McCarthy, John. 1979. Formal problems in Semitic phonology and morphology. Doctoral Dissertation, Department of Linguistics and Philosophy, MIT.

McCarthy, John. 1988. Feature geometry and dependency: a review. *Phonetica* 45:84–108.

McCarthy, John. 2002. *A Thematic Guide to Optimality Theor*y. Cambridge: Cambridge University Press.

McGarrity, Laura W. 2002. On the typological predictions of fixed vs. complementary rankings of stress constraints. In *Online Proceedings of the 2002 Texas Linguistics Societ*y, ed. A. Aguwele and H. Park. University of Texas, Austin: Texas Linguistic Forum.

Mikheev, Andrei. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics* 23:405–423.

Mohri, Mehryar. 2002. Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics* 7:321–350.

Myers, Scott. 2002. Gaps in factorial typology: the case of voicing in consonant clusters. URL http://uts.cc.utexas.edu/ smyers/voicing.pdf.

O'Connor, J. D. and J. L. M. Trim. 1953. Vowel, consonant, and syllable—a phonological definition. *Word* 9:103-122

Ohala, John J. 1981. The listener as the source of sound change. In *Papers from the Parasession on Language and Behavio*r, ed. Carrie S. Masek, Roberta. A. Hendrick, and Mary Frances Miller, 178–203.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological reality of morpheme structure constraints. In *Experimental phonolog*y, ed. John J. Ohala and Jeri J. Jaeger, 239–252. San Diego, CA: Academic Press.

Pater, J., and A.-M. Tessier. 2003. Phonotactic knowledge and the acquisition of alternations. In *Proceedings of the 15th International Congress on Phonetic Science*s, 1777–1180. Barcelona.

Pater, Joe, and Andries Coetzee. 2006. Lexically ranked OCP-Place constraints in Muna. Ms. under review.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. In press. The acquisition of allophonic rules: statistical learning with linguistic constraints. To appear in *Cognition*.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology II*I, ed. Patricia Keating, 168–188. Cambridge University Press.

Pierrehumbert, Janet. 2001a. Stochastic phonology. *GLOT 5* 1–13.

Pierrehumbert, Janet. 2001b. Why phonological constraints are so coarse-grained. In *SWAP special issue, Language and Cognitive Process*es, ed. James McQueen and Anne Cutler, volume 16, 691–698.

Pierrehumbert, Janet. 2006. Incremental learning of the phonological grammar. Talk presented at the 80th Annual Meeting of the Linguistic Society of America.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1992. *Numerical Recipes in C: The Art of Scientific Computin*g. Cambridge: Cambridge University Press.

Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14:19–100.

Prince, Alan. 1985. Improving tree theory. In *Proceedings of the Berkeley Linguistics Society 1*1, 471–490. Berkeley, CA: Berkeley Linguistics Society.

Prince, Alan. 1990. Quantitative consequences of rhythmic organization. In *CLS 26-II, Papers from the Parasession on the Syllable in Phonetics and Phonology*, ed. Karen Deaton, Manuela Noske, and Michael Ziolkowski, 355–398. Chicago Linguistic Society.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Cambridge, MA: Blackwell. [Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993].

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing Priorities: Constraints in Phonological Acquisitio*n, ed. René Kager, Joe Pater, and Wim Zonneveld, 245– 291. Cambridge: Cambridge University Press.

Rastle, Kathleen, Jonathan Harrington, and Max Coltheart. 2002. The ARC Nonword Database. *Quarterly Journal of Experimental Psychology: Section A* 55:1339–1362.

Riggle, Jason. 1999. Relational markedness in Bantu vowel harmony. In *Phonology at Santa Cruz*, ed. Adam Ussishkin, Nathan Sanders, and Dylan Herrick, volume 6, 57–70. Academic Press.

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

Rose, Sharon, and Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80:475–531.

Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language 1996* 10:187– 228. [Long version: Carnegie Mellon Tech. Rep. CMU-CS-94-138].

Ross, John R. 1972. The category squish: Endstation Hauptwort. In *Chicago Linguistic Society* 8, 316–328. University of Chicago.

Scholes, Robert. 1966. *Phonotactic Grammaticality*. The Hague: Mouton.
Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality and Linguistic Methodology*. Chicago: University of Chicago Press.
Selkirk, Elisabeth O. 1980a. Prosodic domains in phonology: Sanskrit revisited. In *Juncture*, ed. Mark Aronoff and Mary Louise Kean, 107-129. Saratoga, CA: Anma Libri.
Selkirk, Elisabeth O. 1980b. The role of prosodic categories in English word stress. *Linguistic Inquiry* 11:563–605.
Selkirk, Elisabeth. 1982. The syllable. In *The structure of phonological representations (part II)*, ed. Harry van der Hulst and Norval Smith, 337–383. Dordrecht: Foris.
Sherer, Tim. 1994. Prosodic Phonotactics. Doctoral Dissertation, University of Massachusetts, Amherst.
Smith, Jennifer. 2001. Lexical category and phonological contrast. In *Papers in experimental and theoretical linguistics 6: Workshop on the Lexicon in Phonetics and Phonology*, ed. Robert Kirchner, Joe Pater, and Wolf Wikely, 61–72. Edmonton: University of Alberta.
Smolensky, Paul. 1986. Information processing in dynamical systems: foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. David E. Rumelhart, James L. McClelland, and the PDP Research Group, volume 1, 194–281. Cambridge, MA: MIT Press/Bradford Books.
Sorace, Antonella, and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.
Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 43:393–436.
Steriade, Donca. 1987. Redundant values. In *CLS Parasession on Autosegmental and Metrical phonology*, ed. A. Bosch, B. Need, and E. Schiller, 339–362. Chicago: Chicago Linguistic Society.
Steriade, Donca. 1995. Underspecification and markedness. In *The Handbook of Phonological Theory*, ed. John Goldsmith, 114–174. Oxford: Blackwell.
Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. In *Proceedings of the 1998 Linguistics and Phonetics Conference*, ed. Osamu Fujimura, Brian Joseph, and B. Palek, 205–245. Prague: The Karolinum Press.
Steriade, Donca. 2001a. Directional asymmetries in place assimilation: A perceptual account. In *The Role of Speech Perception in Phonology*, ed. E. Hume and K. Johnson, 219–250. San Diego: Academic Press.
Steriade, Donca. 2001b. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. Ms., MIT.
Tesar, Bruce. 2004. Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry* 35:219–253.
Tesar, Bruce, and Alan Prince. to appear. Using phonotactics to learn phonological alternations. In *Proceedings of the Thirty-Ninth Conference of the Chicago Linguistics Society, vol. II: The panels*. Chicago: Chicago Linguistic Society.
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, MA: MIT Press.
Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, ed. Michael B. Broe and Janet Pierrehumbert, 269–282. Cambridge: Cambridge University Press.
Vergnaud, Jean-Roger. 1977. Formal properties of phonological rules. In *Basic Problems in Methodology and Linguistics*, ed. R. Butts and J. Hintikka. Amsterdam: Reidel.
Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce, and David Kemmerer. 1997. Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech* 40:47–62.
Whorf, Benjamin L. 1940. Linguistics as an exact science. *Technology Review* 43: 61-63, 80-83.
Wilson, Colin. to appear. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* .