

---

# On the Role of Locality in Learning Stress Patterns\*

Jeffrey Heinz

November 28, 2007

## Abstract

This paper presents a previously unnoticed universal property of stress patterns in the world's languages: they are, for small neighborhoods, neighborhood-distinct. Neighborhood-distinctness is a locality condition defined in automata-theoretic terms. This universal is established by examining stress patterns contained in two typological studies, Bailey (1995) and Gordon (2002). Strikingly, many logically possible—but unattested—patterns do not have this property. Not only does neighborhood-distinctness unite the attested patterns in a non-trivial way, it also naturally provides an inductive principle allowing learners to generalise from limited data. A learning algorithm is presented which generalises by failing to distinguish same-neighborhood environments perceived in the learner's linguistic input—hence learning neighborhood-distinct patterns—as well as almost every stress pattern in the typology. In this way, this work lends support to the idea that properties of the learner can explain certain properties of the attested typology, an idea not straightforwardly available in Optimality-theoretic and Principle and Parameter frameworks.

---

\*Much of this research was performed with the support of a UCLA Dissertation Year Fellowship. I also would like to thank Rachel Schwartz and Stephen Tran for their efforts in helping to collect secondary stress patterns. Portions of this work were presented in 2006 at SIGPHON and at NELS, and in 2007 at the University of Delaware, the University of California, Los Angeles, and Oakland University. I would like to thank those audiences for their comments and suggestions. I also especially thank Edward Stabler and Kie Zuraw for invaluable discussion, as well as Bruce Hayes, Gregory Kobele, Stott Parker, Katya Pertsova, Sarah van Wagnenen, Colin Wilson, and the participants of the 2007 fall seminar on learning theory at the University of Delaware.

Many of the things you can count, don't count. Many of the things you  
can't count, really count. Albert Einstein

## 1 Introduction

This paper presents a previously unnoticed universal property of the stress patterns in the world's languages: they are, for small neighborhoods, neighborhood-distinct (these terms are defined in §4 below). This universal is established by examining the stress patterns contained in two recent typological studies, Bailey (1995) and Gordon (2002). This universal is interesting for three reasons. First it speaks directly to the notion of locality in phonology. Second, many logically possible—but unattested—patterns do not have this property. In other words, despite the extensive variation present in the stress patterns included in the typological studies of Bailey (1995) and Gordon (2002), the property of neighborhood-distinctness unites the attested patterns in a non-trivial way. Third, neighborhood-distinctness naturally provides an inductive principle learners can use to generalise correctly from limited data.

### 1.1 Two Hypotheses

This paper is in two parts. The first part motivates representing stress rules (and phonotactic patterns in general) with regular sets (i.e. finite-state automata, see also Idsardi (2005)), motivates and defines neighborhood-distinctness as a locality condition in phonology, and applies the definition to stress patterns in Bailey's (1995) and Gordon's (2002) typologies to reveal the universality of this property. This result constitutes this paper's first hypothesis:

- (1) All phonotactic patterns are neighborhood-distinct.

The hypothesis is stated in terms of phonotactic patterns, as opposed to stress patterns, because in languages in which the stress pattern is predictable in some (given) domain, the pattern can be thought of as a constraint on the well-formedness of phonological strings within that domain. Under this view, such stress patterns become a class of phonotactic patterns, and in the interests of developing strong testable hypotheses, I frame the hypothesis in (1) as strongly as possible.

The second part addresses the significance of this finding, and how might it be addressed in a theory of phonology. One possibility that is discussed is to place a condition on CON, that component of an Optimality-theoretic grammar (Prince & Smolensky 1993, 2004) in which universal phonological constraints lie (cf. Eisner (1997b), McCarthy (2003)). This paper however explores another possibility: that patterns are neighborhood-distinct because the learner itself is unable to distinguish between the same-neighborhood environments present in the learner's linguistic input.

The inability on the part of the learner to make such a distinction would explain why the attested patterns are neighborhood-distinct—since the learner could only acquire patterns in which the neighborhood of every environment is distinct. As it turns out, the proposed learner—called the Forward Backward Neighborhood Learner—is unable to learn every neighborhood-distinct pattern; however, it does succeed on 100 of the 109

patterns in the typology.<sup>1</sup> Although the results are not perfect, they are comparable to the results of previous learners (Dresher & Kaye 1990, Gupta & Touretzky 1991, Goldsmith 1994, Tesar 1998, Tesar & Smolensky 2000). Furthermore, as is discussed, since the stress rules obtained by the learner in these ‘failure’ cases do not differ greatly from the rules proposed by phonologists, there is an open question as to whether further empirical work on these languages vindicates the learning proposal. In other words the learning algorithm introduced here leads to a second hypothesis:

- (2) Phonotactic patterns are in the range of the Forward Backward Neighborhood Learner.

It is not known which of Hypotheses 1 and 2 is the stronger (i.e. more restrictive hypothesis).

Hypotheses 1 and 2 are claims about the nature of locality in phonology. I emphasise that neither claims that locality is the only relevant factor in phonotactic patterns or phonotactic learning. There are clearly many relevant factors in learning phonotactic patterns: articulatory, perceptual, sociolinguistic, etc. In particular, the learning study presented later can be understood as an investigation into the contribution locality can make to learning stress patterns. Factoring the learning problem in this way—i.e. investigating the contributions individual factors can make to the learning process—helps us understand which factors are necessary, sufficient or irrelevant, and ultimately what the cumulative effects of different combinations of factors yield.

Together Hypotheses 1 and 2 constitute a step towards developing a learning-based theory of grammar, of which one goal might be construed as explaining some natural language patterns with properties (or biases) of the learner. In this respect, this work shares the same goal as other recent proposals Wilson (2006), Moreton (2007), Finley & Badecker (2007) in trying to determine to what extent biases of the learner can explain typological facts.<sup>2</sup> Although this work differs from these others in its focus—those works investigate how phonetic or substantive bias affect the way learners generalise, whereas here the focus is on a particular formulation of locality—the goal is the same: learning-based explanations of phonological typology. It is my hope that this paper leads not only to further investigation into the formal properties of these classes of patterns, but also guides future empirical work—typological, experimental, and descriptive.

## 1.2 Approaches to Learning in Phonology

This approach to the learning problem—where the generalization strategy of the learner directly relates to inherent properties of the hypothesis space—is different from the ones taken in the P&P and OT frameworks. In those approaches, the proposed learning mechanisms operate over an additional layer of structure provided by the framework which is independent of any inherent properties of the hypothesis space.

It is easy to see that this is true by recognizing that the learning algorithms that have been proposed for P&P and OT grammars are essentially the same no matter

---

<sup>1</sup>Code is available on the author’s website for running the learner.

<sup>2</sup>See also related discussion in Stabler (2007).

what particular set of constraints or parameters is adopted—in other words, no matter what the predicted typology is. If Universal Grammar (UG) carved out some other hypothesis space, the proposed learning algorithms would not have to change. This is not controversial. Indeed, Tesar and Smolensky (2000:7-8) make it quite plain:

OT is a theory of UG that provides sufficient structure at the level of the grammatical framework itself to allow general but grammatically informed learning algorithms to be defined. . . Yet the structure that makes these algorithms possible is not the structure of a theory of stress, nor a theory of phonology: it is the structure defining any OT grammar. . .

Dresher (1999) makes the same point, also with respect to learners proposed in the P&P framework:

. . . at the most general level. . . the learning algorithm is independent of the content of the grammar. . . for example, . . . it makes no difference to the [Triggering Learning Algorithm (Gibson & Wexler 1994)] what the content of a parameter is: the same chart serves for syntactic word order parameters as for parameters of metrical theory, or even for nonlinguistic parameters.

In other words, in the P&P and OT frameworks, the proposed learning mechanisms operate over the structure provided by the framework, and not any inherent structure that may exist in the hypothesis space itself.<sup>3</sup>

The approach taken here also differs from the recent explosion of work in probabilistic-based learning: approaches based on OT, (Boersma 1997, Boersma & Hayes 2001, Tessier 2006), minimum distance length (Ellison 1994b), Bayes law (Tenenbaum 1999, Goldwater 2006), maximum entropy (Goldwater and Johnson 2003, Hayes and Wilson to appear) and approaches inspired by Darwinian-like processes (Clark 1992, Yang 2000, Martin 2007). These models, whose advantages include being robust in the presence of noise and being capable of handling variation, are primarily methods which effectively search a given hypothesis space. Thus, these learners are *structured* probabilistic models. Again, this is not controversial. For example, Goldwater (2006:19) explains that ‘the focus of the Bayesian approach to cognitive modeling is on the probabilistic model itself, rather than on the specifics of the inference procedure.’ Yang (2000:22) describes one of the ‘virtues’ of his approach this way: ‘UG provides the hypothesis space and statistical learning provides the mechanism.’ In other words, if UG provided some other hypothesis space, there would be no need to alter the statistical learning mechanism.

On the other hand, one focus of this paper (Hypothesis 2) is on the shape, or structure, of the hypothesis space itself, as a consequence of the inference procedure, as opposed to the search that takes place within it. The generalization strategy of the learner is what determines the hypothesis space.

---

<sup>3</sup>Dresher (1999:28) draws a distinction between the Triggering Learning Algorithm (TLA) and the ordered cue learning model of Dresher & Kaye (1990), explaining that ‘Cues must be appropriate to their parameters in the sense that the cue must reflect a fundamental property of the parameter, rather than being fortuitously related to it.’ This is a step in the right direction, but neither Dresher & Kaye (1990) nor Dresher (1999) offer a precise explanation of what a ‘fundamental property’ of a parameter would look like, or what properties of an associated cue make it appropriate. Thus it is not exactly clear how different the ordered cue based learner is from the TLA in this respect (see Gillis *et al.* (1995) for further discussion on this point).

Finally, it is useful to point out that the observation that most learning proposals in phonology do not make use of properties inherent in the hypothesis space itself is just that—an observation. It certainly does not constitute an argument against such proposals because it is logically possible that human learners do make use of the additional structure afforded by P&P or OT frameworks. Here I only wish to point out the idea that the hypothesis space as a consequence of the learner—the idea that properties of the learner determine properties of the typology—is a natural one that has not, to my knowledge, been explored in models of phonological acquisition.

### 1.3 Organization

The paper is organised as follows. §2 describes the stress typology which constitutes the empirical data used in this study. §3 motivates representing phonotactic patterns with regular sets (i.e. finite state machines). §4 defines neighborhood-distinctness, makes clear its relevance to locality in phonology, and applies it to the stress patterns in the typology. §5 discusses how the universality of neighborhood-distinctness should be handled by a theory of phonology. §6 introduces the learning framework, defines the Forward Backward Learner (FBL) and gives the results of the study. §7 analyzes and interprets these results, and shows how the learner can be modified to be made incremental. §8 compares the FBL to other learning algorithms that have been evaluated in the domain of stress. §9 summarises the results and suggests future research directions.

There are two appendices included with this paper. The first enumerates the distinct stress patterns in the typology, describes the patterns, and shows the results of the learning algorithm. The second appendix includes a proof of the convergence of the incremental version of the learner.

## 2 The Stress Typology

The choice to study stress systems was made primarily because they are a well-studied part of phonological theory and the attested typology is well-established (Hyman 1977, Hayes 1981, Prince 1983, Halle & Vergnaud 1987, Idsardi 1992, Bailey 1995, Hayes 1995, Hyde 2002, Gordon 2002). Additionally, learning of stress systems has been approached before, e.g. Dresher & Kaye (1990), Goldsmith (1994), Gillis *et al.* (1995), Gupta & Touretzky (1994), Tesar (1998), Tesar & Smolensky (2000), making it possible to compare learners and results.

### 2.1 Summary of the Typology

Combining Bailey’s (1995) and Gordon’s (2002) stress typologies yields a typology of 423 languages, exhibiting 109 distinct stress patterns, representing over 70 language families.<sup>4</sup> These patterns are broadly categorised into three groups by primary stress

---

<sup>4</sup>The stress database *StressTyp* currently maintained by Harry van der Hulst and Rob Goedemans did not become available online until after this project was underway. Many of the same languages in Bailey (1995) and Gordon (2002) are included in *StressTyp*, but *StressTyp* includes more languages

placement: quantity-insensitive, quantity-sensitive bounded, and quantity-sensitive unbounded. The appendix organises the stress patterns in the typology into these three categories and shows the extensive variation documented in those studies. Below for the sake of completeness, I briefly review these categories and some of the variation, though undoubtedly most of the discussion is familiar to anyone with anything greater than a passing interest in stress patterns. For further details, the reader is referred to Bailey (1995), Gordon (2002), Hayes (1995), and the primary source references therein.

Quantity-insensitive (QI) stress patterns, extensively reviewed in Gordon (2002), are those in which stating the stress rule need not refer to the quantity, or weight, of the syllables. A review of the typology reveals 319 languages to be quantity-insensitive. These 319 languages exhibit 39 distinct stress patterns. These patterns can be divided into four kinds: single, dual, binary and ternary systems (Gordon 2002). Single stress systems have a single stressed syllable in each word. Dual stress systems have at most two stressed syllables in each word. Binary and ternary systems have no fixed upper bound on the number of stressed syllables in a word and place stress on every second or third syllable, respectively. All QI systems are bounded—that is primary stress falls within some window of the word edge. No other kind of QI stress system is attested.

Quantity-sensitive (QS) stress systems are unlike QI stress systems in that stress placement is predictable only if reference is made to syllable types. Because syllable distinctions are usually describable in terms of the quantity, or weight, of a syllable, these such patterns are called quantity-sensitive.<sup>5</sup> Because Bailey’s (1995) typology only describes the placement of primary stress, secondary stress information on each of those languages was collected from primary sources when available.<sup>6</sup> The resulting typology includes 44 patterns which have quantity-sensitive bounded patterns.

Like the QI patterns, QS bounded patterns can be subdivided in single, dual, binary, ternary, and ‘multiple’ types (‘multiple’ is defined momentarily). Because of the weight distinction, each of these subtypes shows extensive variation. There are 58 QS bounded languages in the typology, exhibiting 44 stress patterns.

Some languages assign primary stress like the single systems described above, and place secondary stress only on heavy syllables, e.g. Cambodian. These patterns I call ‘multiple’ QS patterns. They are similar to binary and ternary patterns since there is no clear principled upper limit on how many syllables in a word can receive stress. But they differ from binary and ternary patterns in that any number of unstressed syllables can occur between stresses. They are included here with QS bounded systems because the location of primary stress is bounded.

Unlike QI patterns, not all QS systems are bounded. QS unbounded stress systems place no limits on the distances between primary stress and word edges, as primary stress usually falls on the leftmost (or rightmost) heavy syllable from the left (or right) word edge. These are the four basic groupings, but again variation exists within each of these subgroups. There are 45 QS unbounded languages in the typology, exhibiting 26 stress patterns.

---

than the ones in the Bailey (1995) and Gordon (2002) combined. Their database is available here <http://stresstyp.leidenuniv.nl/>. Also see Goedemans *et al.* (1996).

<sup>5</sup>Another proposed dimension along which syllable type can be distinguished is prominence (Hayes 1995). See also Crowhurst & Michael (2005).

<sup>6</sup>This was done with the assistance of [names] undergraduate students at [author’s institution].

## 2.2 Phonotactic Restrictions

In addition to the different stress assignment rules described below, there is additional variation that is relevant to a learner of stress patterns. Some languages place additional restrictions on what strings of syllables are well-formed. Many languages prohibit monosyllabic words, or words consisting of a single light syllable. Other languages require every word to have at least (or at most) one heavy syllable. These phonotactic constraints matter for a learner because words which violate these phonotactic constraints are never present in the learner’s linguistic environment, not even potentially. Therefore, whenever such a restriction was mentioned in a source, it was noted. These restrictions are included in the typology and contribute to the total number of distinct patterns. For example, Alawa (Sharpe 1972) and Mohawk (Michelson 1988) both assign stress to the penultimate syllable, but words in Mohawk are minimally disyllabic, which is not the case as far as I know in Alawa, and so both of these patterns, which differ minimally in this respect, are included in the typology.

## 2.3 Unattested Stress Patterns

Despite the extensive variation recounted above, stress patterns are not arbitrary. There are many, many logically possible ways to assign stress which are unattested. No language places a stress on the fourth syllable from the right (or left) in words four syllables or longer, and on the first (or final) syllable in words three syllables or less. No stress pattern places stress on every fourth or every fifth syllable (cf. binary and ternary patterns above which place stress on every second or third syllable). Moving further afield, languages do not place stress on every  $n$ th syllable, where  $n$  is a prime number, nor on every  $n$ th syllable where  $n$  is equal to some prime number minus one. When we consider the myriad of logically possible ways stress can be assigned, the attested variation appears quite constrained. This is not a new observation—virtually every previous researcher essentially makes the same point. I emphasise it only to beg the question what property or properties the attested patterns share which separates them from these unattested patterns.

## 3 Representing Phonotactic Patterns with Regular Sets

Regular sets are sets of strings—i.e. formal languages or patterns—that can be generated by finite state acceptors.<sup>7</sup> There are several excellent introductions to finite state machines, including Hopcroft *et al.* (2001), Sipser (1997) to which I refer interested readers. I find it useful to represent phonotactic patterns as regular sets, and consequently to consider finite state representations of the grammars which generate these patterns.<sup>8</sup> There are four reasons for this.

<sup>7</sup>See Kracht (2003) Chapter 2 for additional characterizations of regular sets.

<sup>8</sup>It is equivalent to speak of the acceptor as recognizing only those words which obey the stress rule represented by the acceptor.

The first is that virtually all phonotactic patterns—and all known stress patterns—are describable as regular sets. To see this, consider first that most phonological processes are described as functions which map underlying forms to surface forms. This is true both in the rule-based formalisms associated with Sound Pattern of English (SPE) (Chomsky & Halle 1968) and in Optimality Theory (McCarthy & Prince 1993, Prince & Smolensky 2004). It has long been observed that almost all phonological processes are regular (Johnson 1972, Kaplan & Kay 1981, 1994). Specifically this means that the function which maps underlying forms to surface forms is a finite state function, and can be represented with a finite state transducer. For phonological grammars, the range of this function is properly interpreted as set of possible legal surface forms; i.e. well-formed words. Consequently, phonotactic patterns are regular because of the well known fact that the range of a finite state function is a regular set (Hopcroft *et al.* 2001).

As an example, the finite state acceptor in Figure 1 is a representation of the stress pattern of Selkup. Selkup assigns stress according to the ‘Leftmost Heavy Otherwise Rightmost’ (LHOR) stress rule, stated below in (3) (Idsardi 1992, Walker 2000).

- (3) Place stress on the leftmost heavy syllable in the word. If there are no heavy syllables, stress the rightmost syllable.

In finite-state diagrams in this paper, start states are indicated by triangles, and final states with double peripheries. This acceptor meets the minimum requirement for a

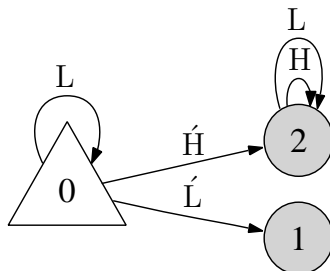


Figure 1: The Leftmost Heavy Otherwise Rightmost (LHOR) stress pattern.

phonotactic grammar—a device that at least answers Yes or No when asked if some word is possible (Chomsky 1957, Chomsky & Halle 1968, Halle 1978). It is easy to verify that every word this grammar accepts obeys the LHOR stress rule and every word it rejects disobeys it. (The words in Table 1 show words of four or fewer syllables which obey the LHOR stress rule, and are exactly those recognised/generated by the machine in Figure 1.) In fact, this grammar recognises infinitely many legal words, just like the generative grammars of earlier researchers.

Second, these regular sets can be related to traditional phonological grammars. Johnson (1972) and Kaplan and Kay (1994) have shown how to construct this finite state transducer which maps underlying forms to surface forms from traditional SPE-style rule-based phonological grammars. Similarly, Riggle (2004), building on work by Ellison (1994a), Eisner (1997b), Karttunen (1998), Frank & Satta (1998), Albro (1998, 2005), shows how to construct a finite state transducer with OT grammars. Given that

H́	Ĺ	H́ L	H́ H	L H́
L Ĺ	H́ L L	H́ L H	H́ H L	H́ H H
L H́ L	L H́ H	L L Ĺ	L L H́	L H́ L L
L H́ L H	H́ L L L	H́ L L H	H́ H L L	H́ H L H
L H́ H L	L H́ H H	H́ L H L	H́ L H H	H́ H H L
H́ H H H	L L H́ L	L L H́ H	L L L Ĺ	L L L H́

Table 1: Words of four or fewer syllables which obey the LHOR stress pattern

it is a simple matter to convert a finite state transducer, which maps underlying forms to surface forms, to a finite state acceptor which accepts only the surface forms (i.e. well-formed words) of the language, it is possible to compute a phonotactic acceptor for well-formed words given more traditional SPE or OT grammars.

For example, consider the (different) OT analyses of the LHOR pattern given in Walker (2000) and Baković (2004). If these analyses were encoded in finite-state OT, applying Riggle’s (2004) transducer-construction algorithm would yield the (same) acceptor shown above. In a sense, the acceptor in Figure 1 implicitly embodies all the constraints and the rankings used in those analyses.

Third, a literature exists on learning regular sets. E.g. the class of regular languages is not exactly identifiable in the limit from positive data (Gold 1967, Angluin 1980), but certain subsets of it are (e.g. Angluin (1982)). Thus it becomes possible to ask: What subset of the regular languages delimits the class of possible human phonotactics and do the properties of this class provide inductive principles for learners?

Finally, insights in this domain can be extended if it is determined that more complex types of grammars are needed. For example, Albro (2005) makes restricted extensions to a finite state system in order to handle reduplication. Also if the working assumption that phonotactic constraints are categorical is relaxed, stochastic finite state automata are a natural extension in which gradient well-formedness patterns can now be described (see Hayes & Wilson (2008) for thorough review of the evidence for gradient well-formedness judgments.)

## 4 The Neighborhood-distinct Hypothesis

In this section I define the concept of neighborhood-distinctness and make clear how it is a formulation of locality in phonology.

### 4.1 Locality in Phonology

It is generally agreed that locality is an important feature of phonological grammars. McCarthy and Prince (1986:1) write ‘Consider first the role of counting in grammar. How long may a count run? General considerations of locality, ... suggest that the answer is probably ‘up to two’: a rule may fix on one specified element and examine a structurally adjacent element and no other.’ Similarly, Kenstowicz (1994:597) also refers to ‘... the well-established generalization that linguistic rules do not count beyond two ...’.



Figure 2: Two States with the Same Neighborhood

Within the particular domain of stress, the thinking is no different. In their *Essay on Stress*, Halle and Vergnaud write ‘...it was felt that phonological processes are essentially local and that all cases of nonlocality should derive from universal properties of rule application’ (1987:ix). Also, Hayes (1995:34) writes ‘Metrical theory forms part of a general research program to define the ways in which phonological rules may apply non-locally by characterizing such rules as local with respect to a particular representation.’

Focusing exclusively on the role of locality does not mean other factors are unimportant or irrelevant to systems of stress. The attention given to it here is made for two reasons: (1) see in precisely what way the stress patterns in the etymology of stress are local and (2) to obtain a clear understanding of the contribution this a priori notion of locality can make to learning.

## 4.2 Definition

The key idea is that each state in a finite state acceptor represents some phonological environment (for some related discussion see Riggle (2004)). Given the idea that phonological environments are ‘local’, we can identify each state with its local characteristics. Thus I define the neighborhood of a state as in (4) below (to be revised):

- (4)
1. the set of incoming symbols to the state
  2. the set of outgoing symbols to the state
  3. whether it is a final state or not
  4. whether it is a start state or not

Thus the neighborhood of state can be determined by looking solely at whether or not it is final, whether or not it is a start state, what the set of paths of length one which reach that state is, and what the set of paths of length one which depart that state is. Pictorially, all the information about the neighborhood of a state is found within the state itself, as well as the transitions going into and out of that state. For example, suppose states  $p$  and  $q$  in Figure 2 belong to some larger acceptor. We can decide that states  $p$  and  $q$  have the same neighborhood because they are both nonfinal, nonstart states, and both can be reached by some element of  $\{a, b\}$ , and both can only be exited by observing a member of  $\{c, d\}$ .

The size of the neighborhood can be parameterised by adjusting parts (1) and (2) of the definition in (4). Instead of referring not just to incoming and outgoing symbols—which are really just paths of length one—those definitions can refer to incoming and outgoing paths of lengths  $j$  and  $k$ , respectively. Thus we define the  $j$ - $k$  neighborhood

as:

- (5)
  1. the set of  $j$ -length paths incoming to the state
  2. the set of  $k$ -length paths outgoing to the state
  3. whether it is a final state or not
  4. whether it is a start state or not

It is now possible to define acceptors that are  $j$ - $k$  neighborhood-distinct.

- (6) An acceptor is said to be  **$j$ - $k$  neighborhood-distinct** iff no two states have the same  $j$ - $k$  neighborhood.<sup>9</sup>

The class of neighborhood-distinct languages is defined in (7).

- (7) **The  $j$ - $k$  neighborhood-distinct languages** are those for which there is an acceptor which is  $j$ - $k$  neighborhood-distinct.

When the values of  $j$  and  $k$  are understood from context, I will just write neighborhood-distinct.

### 4.3 Neighborhood-distinct acceptors and languages

Although many different acceptors can recognise exactly the same language or pattern, certain ones are more useful than others. For example, a forward deterministic acceptor with the fewest states for a language is called the language's tail canonical acceptor and typically finite state patterns are represented with this acceptor. I will refer to such acceptors which are  $j$ - $k$  neighborhood-distinct (and the languages they recognise) as *tail-canonically  $j$ - $k$  neighborhood-distinct*. However, another algebraically equivalent choice is a reverse deterministic acceptor with the fewest states for a language, which is called the language's head canonical acceptor (Heinz 2007). It will be useful to refer to head-canonical acceptors which are neighborhood-distinct (and the languages they recognise) as  *$j$ - $k$  head-canonically neighborhood-distinct*. These two acceptors can be computed from any acceptor which recognise a given pattern (Hopcroft *et al.* 2001). Finally, I will refer to patterns which are either tail or head canonically neighborhood distinct simply as  *$j$ - $k$  canonically neighborhood-distinct*.

### 4.4 Properties of Neighborhood-distinct languages

An analysis of neighborhood-distinct languages based on its component parts is begun in Heinz (2007). The results so far show that the class is finite, that the  $j$ - $k$  neighborhood-distinct languages form a structured hierarchy of classes and that it does not have certain closure properties. These are discussed in turn.

1-1 neighborhood-distinctness is restrictive. The neighborhood-distinct languages not only form a proper subset of the regular languages over some alphabet  $\Sigma$ , there are only finitely many of them: all regular languages whose smallest acceptors have more

---

<sup>9</sup>Also, the acceptor must consist only of useful states—i.e. every state must be reachable from some start state, and the final state must be reachable from any state.

than  $2^{2^{|\Sigma|+1}}$  states cannot be 1-1 neighborhood-distinct (since at least two states would have the same neighborhood). Thus most regular languages are *not* 1-1 neighborhood-distinct.<sup>10</sup>

Also, it is obvious that if a pattern is  $j$ - $k$  neighborhood-distinct then it is  $j'$ - $k'$  neighborhood-distinct where the  $j'$ - $k'$  neighborhood is larger (i.e. where  $j' \geq j + 1$  or for some  $k' \geq k + 1$ ). Thus neighborhood-distinct languages form a hierarchy with language expressiveness increasing as one moves up the hierarchy.

Finally, Heinz (2007:chap. 6) shows that the  $j$ - $k$  neighborhood-distinct languages are not closed under union, intersection, or complement. A language class is said to be closed with respect to some operation if application of the operation to one (or more) languages in the class always yields another language in the class. Since languages here are conceived as sets of strings, union, intersection, and complement have their usual meanings. The absence of these properties—in particular intersection—matters, as we will see below.

## 4.5 Universality

For each of the distinct patterns in the typology, I constructed a finite state acceptor such that only those words which obey the language's stress rules are recognised by the acceptor. Note that the words these machines recognise are strings of syllables, not segments. Thus this study abstracts away from the different ways languages determine the relevant quantity of a syllable for the purpose of assigning stress.<sup>11</sup> These machines are available electronically as part of a stress typology database available at the author's website.

When we consider the typology of stress patterns, 97 are tail canonically 1-1 neighborhood-distinct and 105 are head-canonically 1-1 neighborhood distinct. Only two languages are neither tail nor head canonically 1-1 neighborhood-distinct (though they are canonically 2-2 neighborhood-distinct). In other words, 107 of the 109 types of languages in the stress typology are canonically 1-1 neighborhood-distinct. One of these two non-canonically 1-1 neighborhood-distinct stress patterns is provably not 1-1 neighborhood distinct, the pattern of Içuã Tupi (Abrahamson 1968). It remains an open question whether there is some neighborhood-distinct acceptor which recognises the other one, which is the pattern of Hindi as described by Kelkar (1968). Nevertheless canonical 1-1 neighborhood-distinctness is a near universal property of attested stress patterns, and every attested stress pattern is canonically 2-2 neighborhood-distinct.<sup>12</sup>

---

<sup>10</sup>By similar reasoning, one can see that for any particular values of  $j$  and  $k$ , most regular patterns are not  $j$ - $k$  neighborhood-distinct.

<sup>11</sup>Readers interested in these ways are referred to Gordon (2006).

<sup>12</sup>Space does not warrant discussion, but it is worth noting that significant other classes of phonotactic patterns are also 1-1 neighborhood-distinct, including those adjacency patterns describable with trigram grammars (also called locally 3-testable languages in the strict sense (McNaughton & Papert 1971)) and long distance agreement patterns (Heinz 2007).

## 4.6 Discussion

There is a question as to how serious a challenge the two languages which are not canonically 1-1 neighborhood-distinct are serious obstacles to the hypothesis that all phonotactic patterns are canonically 1-1 neighborhood-distinct (Hypothesis 1). If the two stress patterns in question were common, or from languages whose phonology was well-studied and uncontroversial, the challenge to the hypothesis would obviously be more serious. As it is however, we would like to know more about the patterns in the languages themselves.

Unfortunately, in the case of Içuã Tupi, this is likely impossible as Abrahamson (1968:6) notes that the tribe is ‘almost extinct’ with only two families alive at the time of his studies. According to his paper, Içuã Tupi places stress on the penult in words four syllables or fewer and on the antepenult in longer words. In metrical theory, one would say that final syllable extrametricality is invoked in words with five or more syllables, but not invoked in words with four or less syllables. Although his paper devotes only a few lines to the topic of word stress, there are no obvious errors and the description of the pattern is clear as are the illustrative examples. I see little alternative but to accept the pattern as genuine.

Nonetheless, there are other plausible possibilities which would render the Içuã Tupi pattern canonically 1-1 neighborhood-distinct. For example, Abrahamson (1968) makes no mention of secondary stress. The presence of secondary stress can distinguish states (see discussion of Klamath and Seneca in footnote 26 below). For example, if Içuã Tupi also exhibits secondary stress word-finally, then the pattern becomes neighborhood-distinct. Another possibility is whether stress may optionally be placed on the penult or the antepenult in words with five or more syllables. Although it may be unfair to assume this (as we can expect Abrahamson to have noted it), this alteration also makes the pattern neighborhood-distinct.

On the other hand, the stress pattern of Hindi has been the subject of many different proposals and there is little consensus as to what the stress pattern of Hindi actually is (see Hayes (1995) for discussion). However, even if each different description of Hindi were correct (perhaps because speakers belong to different dialectal groups), a small change to Kelkar’s (1968) description renders it neighborhood-distinct. According to Kelkar, Hindi is a QS unbounded system with a three-way quantity distinction with main stress falling on the rightmost (nonfinal) superheavy, or if there are none, on the rightmost (nonfinal) heavy syllable, or in words with all light syllables, the penult. Secondary stresses fall on heavy syllables and alternate light syllables both sides of main stress.<sup>13</sup> Kelkar’s description of the stress patterns, however, rests on words that are only a few syllables in length. In other words, although his description makes clear predictions about how stress falls in longer words, it is far less clear that these predictions are actually correct. If, in words longer than four syllables, lapses were optionally allowed across two adjacent light syllables and final heavy syllables could optionally bear primary stress instead, then this pattern also becomes neighborhood-distinct.

---

<sup>13</sup>This may also well be the most complicated pattern in the typology, as measured by the number of states in its tail and head canonical acceptors: 26 and 32 respectively (cf. Pirahã which has 33 and 18, respectively).

To sum up, it is premature to reject the hypothesis that all patterns are canonically 1-1 neighborhood-distinct because of the counterexamples of Içuã Tupi and Hindi (per Kelkar). The proposed descriptions of these patterns ought to be investigated further if possible to see if they hold up as counterexamples. A small change in the description of the pattern can render it neighborhood-distinct. Finally, note that the hypothesis that all stress patterns are canonically 1-1 neighborhood-distinct is supported by the many established stress patterns in the typology which fall into this class.

## 5 Significance of the Hypothesis

How can a theory of phonology accommodate the universality of neighborhood-distinctness? The issue is raised because neighborhood-distinctness is a constraint on the well-formedness of an aspect of the total grammar—here, the stress domain. One logical possibility is that neighborhood-distinctness is just an epiphenomenon of a particular set of parameters or constraints that are motivated on independent grounds. Another possibility—the one pursued here—is that it is not an accident that proposed theories all conspire to predict typologies wherein all patterns are neighborhood-distinct.

In Optimality Theory, one approach to account for this systematic gap might be to require that individual constraints be neighborhood-distinct (cf. (Eisner 1997b, McCarthy 2003)). Recall that under many finite-state implementations of OT, a constraint can be represented as a finite-state transducer (Eisner 1997a, Frank & Satta 1998, Riggle 2004, Albro 2005). Thus this approach requires that the notion of neighborhood-distinctness be translated from acceptors to transducers, something which can be done reasonably, though certain details will have to be worked out.<sup>14</sup>

This proposal is interesting and if pursued further, the following remarks apply. First, most phonological constraints are neighborhood-distinct. The most obvious constraints which run afoul for not being neighborhood-distinct are the ALIGN constraints, which have already been shown to be problematic on other grounds (Eisner 1997b, McCarthy 2003). When computing the typology, having only neighborhood-distinct constraints would have the desirable effect that many of the unattested patterns—like those describable with feet of size four or more—are not found within the typology.

However, since the transducer construction algorithm intersects the individual constraints and closure under intersection is not a property of neighborhood-distinct languages, simply ensuring the individual constraints are neighborhood-distinct does not guarantee the neighborhood-distinctness observed in the stress domain observed above. It is a worthwhile endeavor to determine (1) whether concrete proposals of OT constraints from such a restricted CON predict a typology where every pattern is neighborhood-distinct, (2) what properties are present which ensure this outcome, (3) if not, whether the Içuã Tupi and Hindi patterns can be explained in this way, and so on. To sum up, it appears restricting CON to include only neighborhood-distinct constraints can help to explain the neighborhood-distinctness of stress patterns, but there

---

<sup>14</sup>For example, in Riggle’s (2004) framework, where transitions in a machine are marked with an input symbol, an output symbol, and a violation vector, we may decide to count the symbols, but not the vector as part of the neighborhood.

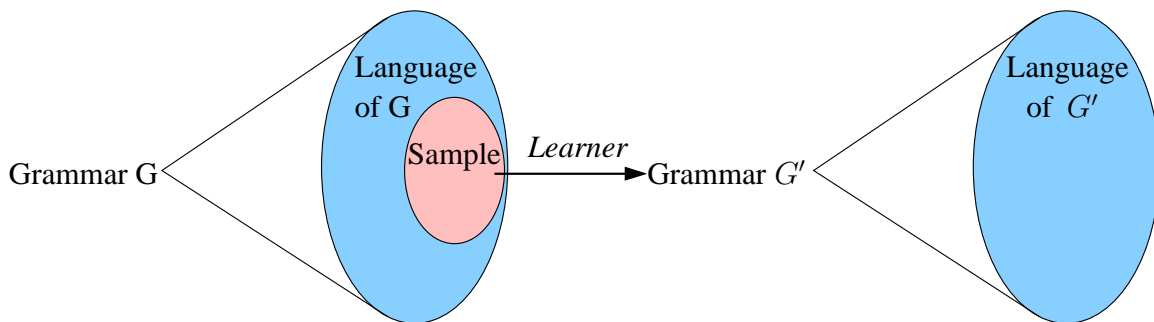
are plenty of unanswered questions and pursuing them may prove a fruitful avenue of research.

This paper pursues another explanation of the neighborhood-distinctness of stress patterns. Namely, stress patterns are neighborhood-distinct because the learner itself is unable to distinguish between the same-neighborhood environments observed in its linguistic environment. The idea underlying this approach is that it is impossible to separate the learner from the hypothesis space because they stand in a natural, intimate relationship: the hypothesis space is the range of the learning function.

## 6 Learning Neighborhood-distinct patterns

### 6.1 The Learning Framework

It is useful to make the learning framework, schematised in Figure 3 explicit. The idea is that language the learner is trying to learn is generated from some grammar  $G$ . The learner, however, does not hear every element of the language (as it is infinite in size), but only some small finite sample. The learner is a function which maps finite samples to grammars. The central question we are interested in is: What is the



learner such that the Language of  $G' = \text{Language of } G$ ? This paradigm is known as exact identification in the limit from positive data Gold (1967). A learner successfully learns language  $L$  if, upon being presented with ever larger finite samples from  $L$ , the grammars returned by the learning function converge to one which generates exactly  $L$ . The class of languages for which a learner can do this is said to be identifiable in the limit from positive data (Niyogi 2006, Nowak *et al.* 2002, Jain *et al.* 1999, Osherson *et al.* 1986).

One key result in framing the learning problem this way is that it is known that learning cannot take place unless the hypothesis space is restricted (Gold 1967, Piattelli-Palmarini 1980, Angluin 1980).<sup>15</sup> In other words,  $G'$  is not drawn from an unrestricted set of possible grammars. The hypotheses available to the learner ultimately determine

---

<sup>15</sup>In particular, the class of all regular sets cannot be exactly identified in the limit from positive data. It is worth noting that this result holds even in other learning frameworks with different criteria for success, e.g. the Probably Approximately Correct framework (Valiant 1984, Blumer *et al.* 1989).

the kinds of generalizations made and the range of possible natural language patterns. Under this perspective, (UG) is this set of available hypotheses.

Given that neighborhood-distinct patterns are a restricted class of languages, it is natural to ask whether there is any learning function which can learn them. Because there are only finitely many neighborhood-distinct languages (see §4.4), there are actually many, many learning functions which can identify this class of patterns in the limit (Jain *et al.* 1999, Osherson *et al.* 1986). However, many of these learners are uninteresting because they make no use of any property of the language class beyond its finiteness.

Therefore, we are interested in a learner for this particular hypothesis space. We might conceive of the learner as making use of the properties defining the space, or more boldly, we might conceive of the properties of the hypothesis space as a consequence of the way the learner works. The Forward Backward Learner presented below is such a learner because it generalises to neighborhood-distinct patterns by its inability to distinguish same-neighborhood states. Note that to the extent this learner succeeds, it explains why stress patterns are neighborhood-distinct.

## 6.2 Overview of the Proposed Learner

Here, I introduce a simple learner which only uses the concept of neighborhood to generalise. The idea is to merge same-neighborhood states in the finite state representation of the input (cf. Biermann & Feldman (1972), Angluin (1982), Oncina *et al.* (1993), Stolcke (1994)). As it turns out, this learner does not identify this class neighborhood-distinct patterns in the limit, though it does succeed for many of the attested stress patterns.

I introduce the learner in two steps for ease of exposition. The first step introduces the Forward Neighborhood Learner first, which succeeds on many, but not all, of the attested patterns. I argue that analysis of the languages which the Forward Neighborhood Learner fails to learn reveals that it is handicapped by the choice of representation of the input. I propose an additional alternative representation of the input and a revised learner called the Forward Backward Neighborhood Learner, which succeeds on many more (but not all) of the attested patterns.

Because the learners are described in automata-theoretic terms, it is possible for future work to provide proven theorems characterizing exactly their behavior. Since those characterizations are still unknown, the proposed learners are evaluated in simulations in the following manner. The input samples presented to the learner in the simulations consist of all words from one to  $n$  syllables which obey the stress pattern.<sup>16</sup> Such a sample is referred to as a sample of size  $n$ . If the acceptor returned by the algorithm did not accept exactly the same language as the target pattern, that was considered a failure. In such a case, the learner was applied to a sample of size  $n + 1$ . If the learning did not occur with samples of size nine, or in some cases eight, I concluded there was no input sample with which the learner would succeed. Note that simulations showed if the learner succeeded for some input sample of size  $n$ , then it also succeeded for any

---

<sup>16</sup>This decision was made primarily for convenience. The alternative is to consider every subset of words from one to  $n$  syllables—an impractical task.

sample of size  $m > n$ . I take this to mean that, for the kinds of samples provided, the learner converges when given the smaller sample of size  $n$ .

Learners were only given words which made the necessary syllable distinctions. For example, QI systems were not given syllables coded for light and heavy, but QS systems which distinguished between syllables of these types were.<sup>17</sup> For example, Table 4 in Appendix A shows that the Forward Backward Learner succeeded in learning the stress pattern of Mam (England 1983) when provided a sample which consisted of all words with one to five syllables. Because this language makes a threeway distinction between syllable types, this means there  $3^5 = 343$  word types that made up the sample.

### 6.3 Prefix Trees

A *prefix tree* is a structured finite state representation of a finite sample. The idea is that each state in the tree corresponds to a unique prefix in the sample. Here ‘prefix’ is not used in the morphological sense of the word, but rather in a mathematical sense meaning ‘initial sequence’. Constructing a prefix tree is a standard algorithm (Angluin 1982). Basically, one can imagine building the tree one word at a time, following an existing path in the tree for as long as possible, and then making a new branch as needed. The prefix tree for the words in Table 1 is shown in Figure 4. The prefix tree accepts only the finitely many forms that have been observed. No generalization has yet taken place. However, even in this simple example, it is possible to see that there is structure in the prefix tree, and that this structure repeats itself. State merging can eliminate structural redundancy, which may result in generalization.

I denote the function which maps some finite sample  $S$  to a prefix tree which accepts exactly  $S$  with  $PT$ . Note that  $PT(S)$  can be computed efficiently in the size of the sample  $S$  (Angluin 1982).

### 6.4 State Merging as Generalization Strategy

The next stage is to generalise by *merging states* in the prefix tree, a process where two states are identified as equivalent and then *merged* (i.e. combined). A key concept behind state merging is that transitions are preserved (Angluin 1982, Hopcroft *et al.* 2001). This is one way in which generalizations may occur—because the post-merged machine accepts everything the pre-merged machine accepts, possibly more.

For example, in Figure 5, Machine B is the machine obtained by merging states 1 and 2 in Machine A. It is necessary to preserve the transitions in Machine A in Machine B. In particular, there must be a transition from state 1 to state 2 in Machine B. There is such a transition, but because states 1 and 2 are the same state in Machine B, the

---

<sup>17</sup>Although this suggests that the learner’s success depends on ‘knowing’ the weight distinction beforehand, this is not so. It is easy to see that if the proposed learners were given words from a QI stress pattern which made some (superficial) light-heavy distinction that the learner would still converge to the correct grammar. This is because additional, unnecessary syllable distinctions do not change the character of the neighborhoods in the target grammars, and the learners operate only on the basis of what distinguishes neighborhoods. In this case, it could be said that the learner discovers that the superficial distinction is unnecessary. The necessary sample size, however, does become larger, which makes the question of what constitutes a *characteristic sample* more pressing (see §9 for further discussion).

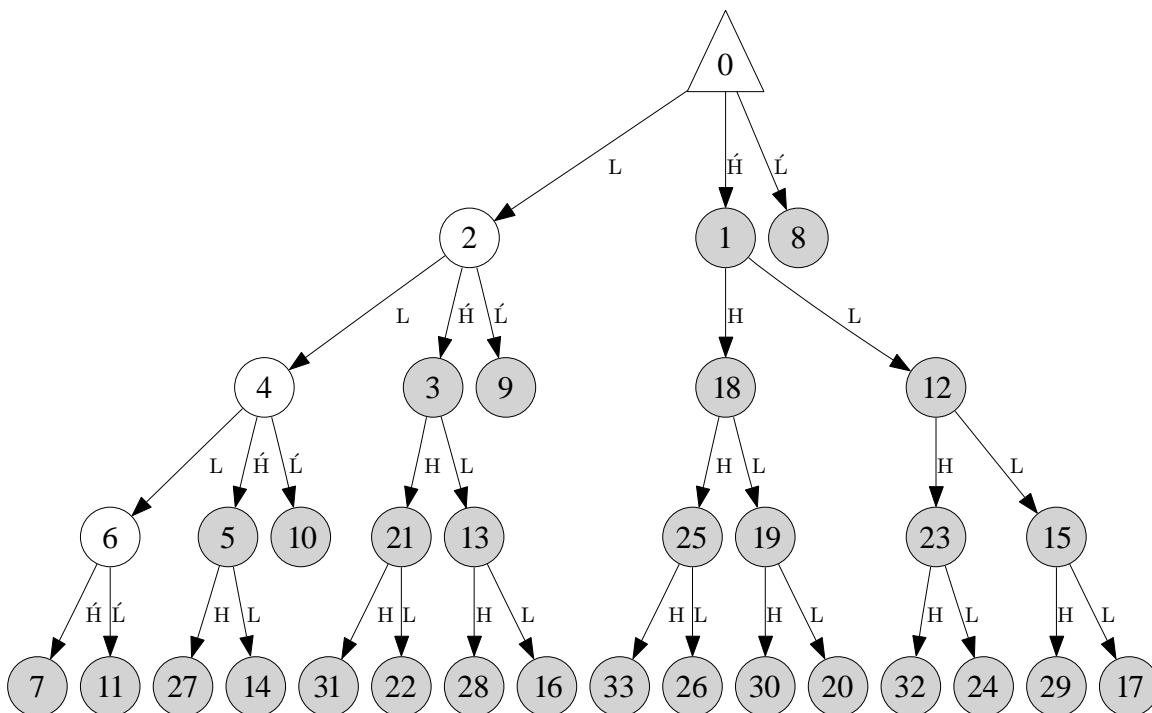


Figure 4: A prefix tree of the LHOR words in Table 1.

transition is now a loop. Whereas Machine A only accepts one word *aaa*, Machine B accepts an infinite number of words *aa, aaa, aaaa, . . .*

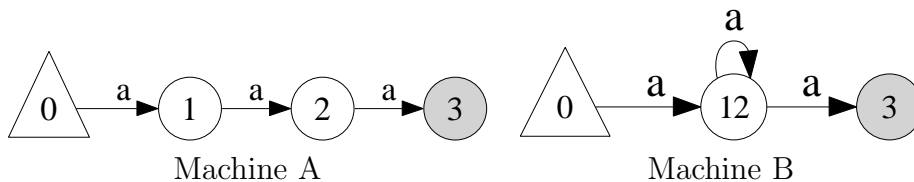


Figure 5: An Example of Generalization by State Merging

Note that the merging process does not specify which states should be merged. It only specifies a mechanism for determining a new machine once it has been decided which states are to be merged. Thus choosing which states are to be merged determines the kinds of generalizations that occur. A merging strategy is thus a generalization strategy. It is an inductive principle, in the sense of (Popper 1959).

One key result is already established (Angluin 1982): Given any canonical acceptor *A* for any regular language *L* and a sufficient sample *S* of *L*—that is, a sample which exercises every transition in *A*—there is some way to merge states in the prefix tree of *S* which returns the acceptor *A*. This result does not tell us how to merge the states for a particular acceptor, it just says that such a way exists. Nonetheless, the result is important because it leaves open the possibility that natural language patterns which form proper subsets of the regular languages (such as stress patterns) have a successful

state merging strategy.<sup>18</sup>

## 6.5 The Forward Neighborhood Learner

The Forward Neighborhood Learner merges states in the prefix tree which have the same 1-1 neighborhood. I denote by  $M_{nd}$  the function which maps a prefix tree  $PT$  to the neighborhood-distinct acceptor obtained by merging all states in  $PT$  with the same 1-1 neighborhood. Note that computing  $M_{nd}$  is efficient in the size of  $PT$ . This is because (1) merging two states is efficient (Hopcroft *et al.* 2001), (2) an algorithm need at most check every pair of distinct states for neighborhood-equivalence to determine if they should be merged, and (3) determining the neighborhood-equivalence of two states is efficient.<sup>19</sup> It is now possible to state precisely the Forward Neighborhood Learner (FNL). The Forward Learner successfully identifies 85 of the 109 pattern types as shown

---

### Algorithm 1 The Forward Neighborhood Learner

---

**Input:** a positive sample  $S$

**Output:** an acceptor  $A$ .

Let  $A = M_{nd}(PT(S))$  and output acceptor  $A$ .

---

in Appendix A.

These results also make clear that the languages in the range of the learning function are not the same as the neighborhood-distinct languages. The two classes of languages clearly overlap, but the Forward Learner does not identify the class of neighborhood-distinct languages in the limit. The Forward Learner does not even identify the tail-canonically neighborhood-distinct languages, falsifying the conjecture made in Heinz (2006) that it does. Nonetheless, the results are promising because the languages for which the Forward learner succeeds cross-cuts the QI, QS bounded, and QS unbounded stress patterns, suggesting the learner is on the right track.<sup>20</sup>

When we examine the languages for which the Forward Learner fails to learn we find that the error is always one of *overgeneralization*. This happens because states are merged which should be kept distinct. Consequently, the grammar returned by the learner accepts a language strictly larger than the target language. This means that there is some word for which the learner’s grammar accepts different stress assignments. This can be construed as optionality—a particular string of syllables can be stressed in one way or another.

Another characteristic that all stress patterns for which the Forward Learner fails share (except Kashmiri<sup>21</sup> is that they are typically analyzed with a metrical unit at the right word edge. Why would such languages be problematic for the Forward Learner?

---

<sup>18</sup>In fact,  $n$ -gram based learning (e.g. see Jurafsky & Martin (2000)) can be described exactly as a particular state-merging procedure (Heinz 2007).

<sup>19</sup>How efficient depends on the representation of the acceptors (i.e. as matrices or as tuples of sets).

<sup>20</sup>In contrast,  $n$ -gram based learners fail to learn the class of unbounded stress patterns with these representations due to the arbitrarily long distances that can occur between stress and a word edge. For further discussion, see Heinz (2007).

<sup>21</sup>According to Walker (2000), the Kashmiri data comes from (Kenstowicz (1993) citing Bhatt (1989)).

One idea is that the prefix tree’s inherent left-right bias fails to distinguish the necessary states, and this occurs more commonly in languages analyzable with a metrical unit at the right word edge. If this were the case, the problem is not with the generalization procedure per se, but rather with the inherent left-right bias of the prefix tree. Below I propose another way the input to the learner can be represented as a finite state acceptor: suffix trees.

## 6.6 Suffix Trees

If the input were represented with a *suffix tree*, then the structure obtained has the reverse bias, a right-to-left bias. Like a prefix tree, a suffix tree is a finite state representation of the input: it accepts exactly the words from which it was built and nothing else. A suffix tree is structured differently from a prefix tree, however, because each state now represents a unique suffix in the sample instead of a prefix. Whereas a prefix tree is forward deterministic, a suffix tree is reverse deterministic. A suffix tree can be constructed in terms of a prefix tree given some sample. This procedure runs as follows: Given a sample of words, build a prefix tree reading each word *in reverse*. Since the resulting prefix tree accepts exactly the reverse of each word in the sample, *reverse* this tree by changing all final states to start states, all start states to final states, and changing the direction of each transition. The resulting acceptor is a suffix tree and accepts exactly the words in the sample.

Figure 6 shows a suffix tree constructed from all words which obey the ‘Leftmost Heavy Otherwise Rightmost’ stress pattern of Selkup up to four syllables in length. Compare the structure of the suffix tree of this representation to the prefix tree shown

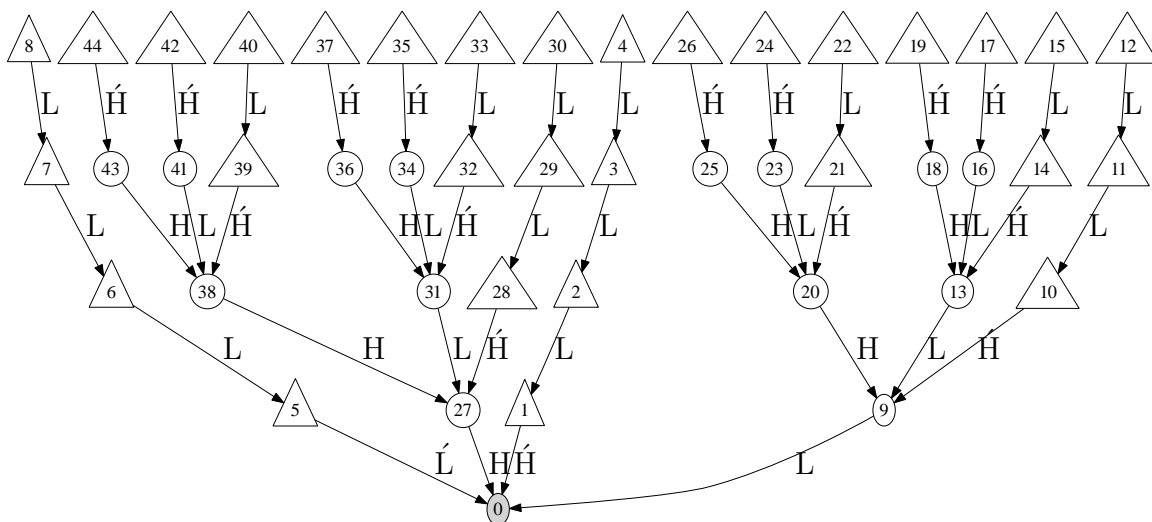


Figure 6: A suffix tree of the LHOR words in Table 1

in Figure 4. Though both representations accept exactly the same finite input, they are not mirror images of each other. The two trees have different structures—though both accept exactly the same (finite) set of words. Because they have different structures, the states in a suffix tree may have different neighborhoods than the states in a prefix

tree. Consequently, the generalizations acquired by merging states with the same neighborhoods may be different.

## 6.7 The Forward Backward Neighborhood Learner

The Forward Backward Neighborhood Learner is very simple. Let  $M_{nd}$  be the function which maps an acceptor to the acceptor obtained by merging same-neighborhood states. Let  $PT$  and  $ST$  denote functions which map a finite sample to the prefix tree and suffix tree, respectively, which accepts exactly the given sample. The learner simply applies the  $M_{nd}$  to the prefix and suffix tree representations of the samples and intersects the results. This learner succeeds on 100 of the 109 patterns (413 of 423

---

### Algorithm 2 The Forward Backward Neighborhood Learner

---

**Input:** a positive sample  $S$

**Output:** an acceptor  $A$ .

Let  $A_1 = M_{nd}(PT(S))$ .

Let  $A_2 = M_{nd}(ST(S))$ .

Let  $A = A_1 \cap A_2$  and output the acceptor  $A$ .<sup>22</sup>

---

languages), a considerable improvement over the Forward Learner. The appendix in §A provides these results, along with those of the Backward Neighborhood Learner (which generalises only by merging same-neighborhood states in the suffix tree).

## 7 Discussion

### 7.1 Basic Reasons Why the Forward Backward Learner Works

The reason the Forward Backward Learner succeeds in more cases than the Forward Learner is simple: intersection keeps the robust generalizations. The robust generalizations are the ones made in *both* the prefix and suffix trees. Overgeneralizations that are made by the Forward Learner are not always made by merging same-neighborhood states in the suffix tree. Consequently, those that are not do not survive the intersection process. Likewise, it is also true that overgeneralizations made by merging same-neighborhood states in the suffix tree are not always made in the prefix tree.

However, the generalization strategy itself—the merging of same-neighborhood states—is the real reason for the algorithm’s success. Consider again the Forward Learner. By merging states with the same neighborhood, the algorithm guarantees that its output is neighborhood-distinct. Similarly, when the same-neighborhood states are merged in the suffix tree, the resulting acceptor is neighborhood distinct. The learner—by merging same-neighborhood states—generalises to neighborhood-distinct

---

<sup>22</sup>Note that intersection ( $\cap$ ) of two acceptors  $A$  and  $B$  results in an acceptor which only accepts words accepted by both  $A$  and  $B$ .

patterns. Thus if people generalise similarly, it explains why nearly all stress patterns are neighborhood-distinct.

There is one caveat, however. As explained in §4.4, the class of neighborhood-distinct languages is not closed under intersection. Thus when the Forward Backward Neighborhood Learner intersects the two acceptors obtained by merging same-neighborhood states in the prefix and suffix trees, the resulting language is not guaranteed to be neighborhood distinct. Little is understood about what additional properties are necessary to ensure that neighborhood-distinctness survives the intersection process. Whatever those properties are, they appear to be in play here. The patterns obtained via the intersection process in the current study produced a tail or head canonically neighborhood-distinct pattern for every pattern in the study (except Ashéninca (Payne 1990)).

## 7.2 Unlearnable Unattested Patterns

It is also interesting to note that most unattested patterns cannot be learned by the Forward Backward Neighborhood Learner. Intuitively, this follows from the fact that neither the Forward Learner nor Backward Learner can ever learn a non-neighborhood-distinct pattern (of which there are infinitely many).

For example, logically possible unattested stress patterns such as those which place stress on every fourth, fifth, sixth, or  $n$ th syllable cannot be learned. To see why, consider the acceptor in Figure 7 which generates the logically possible stress pattern which assigns stress to the initial syllable and then every fourth syllable. The

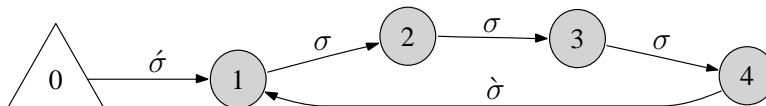


Figure 7: The FSA for a Quaternary Stress Pattern

reason that this pattern cannot be learned by the Forward Backward Neighborhood Learner is because states 2 and 3 have the same neighborhood. It is not possible to write some other acceptor for this language that would not have two states like states 2 and 3 above with the same neighborhood (because the pattern requires exactly three unstressed syllables between stresses). Thus this pattern is not neighborhood-distinct. Consequently neither the Forward Learner nor the Backward Learner could ever arrive at this pattern by merging same-neighborhood states since states 2 and 3 (or more precisely, their corresponding states in the prefix and suffix trees) would always be merged. Furthermore, since this overgeneralization is made by both learners, it survives the intersection process. Thus the result obtained by the Forward Backward Learner is that secondary stresses must occur *at least* two syllables apart. In a sense, the learner fails because it cannot distinguish ‘exactly three’ from ‘at least two.’ Thus, the idea that ‘linguistic rules cannot count past two’ (Kenstowicz 1994:597) is a direct consequence of the way the FBL generalises. Specifically, it is a consequence of generalising by merging same-neighborhood states.

Incidentally, this is the reason for the Einstein quote at the outset: this paper suggests that, in the context of language learning, not being able to count really matters. Learners that cannot count are unable to make certain distinctions, and this will lead to the acquisition of certain types of patterns, but not others. Note it is not immediately obvious that a notion of locality would be even sufficient for learning any stress patterns. This paper has shown that a formal, a priori notion of locality can make a significant contribution to language learning.

Thus the FBL predicts that logically possible stress patterns like the quaternary one above should be at the very least significantly more difficult to learn than ternary or binary patterns. Whether children or adults behave as the FBL predicts—which can conceivably be investigated in artificial language learning experiments—is an open question.

### 7.3 Unlearnable Attested Patterns

In this section, I discuss the nine languages which the Forward Backward Neighborhood Learner failed to learn, which constitute a direct challenge to the viability of Hypothesis 2. Recall that the learner fails if the acceptor obtained does not accept *exactly* the same language.<sup>23</sup> Thus one question to keep in mind throughout this discussion is how different are the patterns obtained by the acceptors to the descriptions found in the typologies.

Again, in every case the learner failed because it overgeneralised. Thus for certain word types, although the grammar obtained by the learner places stress in the correct positions, it can also place stress in other positions.<sup>24</sup> In other words, the learner allows a certain degree of optionality. After reviewing the stress patterns for which the FBL fails, I address these failures below.

The concrete reason why all of these patterns fail is because there are two states which are merged which should not be. In other words, the learner does not distinguish phonological environments where it should have. To make it more concrete than this requires careful examination of the canonical acceptors and the prefix and suffix trees, and space and time prohibit such an extended discussion. Therefore in what follows, I only make a few observations.

Two of the languages for which it fails, Içuã Tupi and Hindi (per Kelkar), are not canonically neighborhood distinct and are discussed in §4.6.

Mingrelian (Klimov 2001) is a neighborhood-distinct pattern described as placing primary stress initially and secondary stress on the antepenult. The FBL fails because it cannot distinguish the sequence of two unstressed syllables at the end of the word from similar sequences in the middle of the word.

The stress patterns of Palestinian Arabic (Brame 1974), Cyrenaican Bedouin Arabic (Mitchell 1975), Negev Bedouin Arabic (Kenstowicz 1983), and Hindi (per Fairbanks (1981)) are all not learnable by this learner, though they are neighborhood-distinct. It is striking that these are precisely the patterns in the typology that have

---

<sup>23</sup>This is a high standard, and one that was relaxed when Valiant (1984) introduced the Probably Approximately Correct Learning framework.

<sup>24</sup>Here by word type, I mean a string of unstressed syllables.

been analyzed with extrametrical feet (Hayes 1995). This suggests that patterns describable with extrametrical feet are beyond the range of the learning function.

Ashéninca (Payne 1990) and Pirahã (Everett 1988) are two other patterns which are neighborhood-distinct but beyond the the range of the learning function. These patterns are well-known prominence systems (Hayes 1995). However, I suspect the reason the Forward Backward Learner fails has less to do with this, than with the fact that both of these languages, like the ones above, can place stress on the third syllable (or the fourth in the case of Ashéninca) from the right edge in particular circumstances. It seems that the Forward Backward Neighborhood learner can learn only some patterns like this (e.g. Walmatjari (Hudson 1978)).

The fact that the FBL fails for stress patterns that are describable with a rule of foot extrametricality (Palestinian Arabic, Cyrenaican Bedouin Arabic, Hindi per Fairbanks, see Hayes (1995)) shows that not all patterns describable in standard metrical theory (Hayes 1995) can be learned by the FBL.<sup>25</sup> The source of this conflict is not well understood except at the most superficial level: the locality conditions imposed by the FBL learner are not met in all patterns describable with extrametrical feet.

Given the hypothesis that the stress patterns are in the range of the FBL learning function, but nine of the stress patterns are not learned by the algorithm, we may conclude that Hypothesis 2 is incorrect. However, such a conclusion is premature for two reasons: the patterns obtained by the learner do not differ greatly from the described patterns and the described patterns for which the learner fails are ones where consensus—if it exists—has formed over a somewhat small data set.

One instructive case comes from Mingrelian (Klimov 2001), which recall places primary stress on the initial and secondary stress on the antepenult. A similar pattern is found in Walmatjari (Hudson 1978), which places stress on the penult in four-syllabled words (presumably to avoid a clash) and *optionally* places stress on the penult or antepenult in longer words. The pattern of Walmatjari is learnable because the states in the acceptor which generate the pattern are made distinct in the suffix tree by the optional penult pattern that occurs in longer words. Interestingly, these are the only two QI dual languages in the typology which place primary stress close to the left word edge and secondary stress on the antepenult. Furthermore, if Mingrelian places secondary stress on the penult in trisyllabic words (or four-syllabled words) to avoid a clash, even optionally, the stress pattern is now learnable (as the relevant states are now distinct in the suffix tree). However, further descriptive research is needed, as Klimov's (2001) study makes no mention of secondary or optional stress, or whether Mingrelian permits a clash in words with four syllables.

---

<sup>25</sup>It is worth asking if there are other stress patterns that are predicted to exist in metrical stress theory (or any of its derivatives) that are either non-neighborhood-distinct or non-learnable by the FBL. This is a project beyond the scope of this paper. One way to proceed might be to see whether the stress patterns generated in a factorial typology of OT constraints are learnable by the FBL. Proposals include Eisner (1998), Tesar (1998), Kager (1999), and Hyde (2002).

One potentially problematic pattern is one where alternating stress occurs on both sides of a primary stress. Different states for the alternating pattern are required to keep track of whether the primary stress has been seen, but the states themselves may have the same neighborhoods. This is like the Yidip pattern, except Yidip is neighborhood-distinct and FBL learnable because of the distinction it maintains between heavy and light syllables.

In other cases, the differences between the acceptor obtained by the FBL and the target pattern are slight. Consider Içuã Tupi, for example: the FBL acceptor predicts that secondary stress may fall optionally on the penult instead of the antepenult in words five syllables or longer. The FBL acceptor obtained for Pirahã can place stress as per the Pirahã pattern or in some words optionally on the final syllable. In Ashéninca, the FBL predicts words ending with a long vowel followed by three syllables with the high front vowel like attested [má:kiriti] ‘type of bee’ could have two pronunciations: [má:kiriti] and [mà:kirítí]. According to Fairbanks, in Hindi stress falls on the initial syllable in disyllabic words. The only overgeneralization made by the FBL is in disyllabic words ending with a superheavy syllable: the initial syllable may be stressed or the superheavy syllable may be stressed (but not both).

The idea that the earlier descriptions are inaccurate does not mean that the actual patterns are completely different from what previous researchers described. In fact, the patterns can differ minimally in interesting ways and even include the same set of words that earlier researchers used to develop their own hypotheses. The two ways that I am suggesting here are (1) in certain words, there will be optionality and (2) in languages currently described as lacking secondary stress, there may in fact be secondary stress. Because theory helps direct the course of investigation, it is plausible that these might be overlooked (or in the case of secondary stress, difficult to detect) in earlier hypothesis formation.<sup>26</sup>

Not all the overgeneralizations made by the learner may be as plausible as the above discussion might suggest. For example, the only overgeneralization made by the FBL when learning Kelkar’s description of Hindi occur in words five syllables or longer. However, some of the optionally acceptable forms include no primary stress (instead secondary stress occurs everywhere expected, including where primary stress should fall). Descriptions of Negev Bedouin Arabic say that if the final syllable is not superheavy and the penult is heavy, stress falls on the penult. However, the acceptor obtained by the FBL accepts words with stress on the penult or final syllable when the last two syllables are heavy (but not both). It is reasonable to expect that this kind of optionality would have been noted by earlier researchers, and therefore the inability of the FBL to learn it is significant. The overgeneralizations in Palestinian Arabic involve only words at least four syllables in length whose penult and antepenult are light, and in Cyrenaican Bedouin Arabic the overgeneralization occurs in certain words three syllables or greater. In these last cases, the amount of optionality appears greater than what optional processes in phonology admit. Nonetheless, a careful review of the descriptions is warranted as is developing a deeper understanding of the nature of the FBL and neighborhood-distinctness.

Finally, it is worth noting that if the FBL was modified so that it only merged states with the same 2-2 neighborhoods, then in fact all overgeneralization is eliminated as

---

<sup>26</sup> We might also expect then that removing secondary stress from some attested patterns may also have an effect. In fact, this makes some learnable patterns unlearnable. For example, it was discovered that if secondary stress is excluded from the grammars of Klamath (Barker 1963, 1964, Hammond 1986, Hayes 1995) and Seneca, (Chafe 1977, Stowell 1979, Prince 1983, Hayes 1995) then the Forward Backward Neighborhood Learner fails to learn these grammars. It fails because, in the actual grammars of Klamath and Seneca, the presence of secondary stress *distinguishes* the neighborhoods of certain states of the prefix and/or suffix trees.

every pattern in the typology can be identified exactly in the limit from positive data.<sup>27</sup>

## 7.4 Learnable Unattested Patterns

The Forward Backward Neighborhood Learner can also learn unattested patterns that are unnatural and/or not present in the typology. In such cases, it is important to remember that the learner developed here only examines the contribution that locality can make to learning. We have seen that this contribution is significant—in fact sufficient for learning to occur—but in no way should we expect it to be the only factor in learning stress patterns, or the only factor which plays a role in determining the typology of human phonotactic patterns.

For example, consider the logically possible stress pattern ‘Leftmost Light Otherwise Rightmost’ (LLOR). Whether or not humans can learn such a pattern is an open question, and to my knowledge there is no experimental evidence bearing on it. However, even if it were shown that LLOR is more difficult to learn than the more natural ‘Leftmost Heavy Otherwise Rightmost’ pattern (suggesting the gap in the typology is systematic and not accidental), the fact is plausibly due to considerations separate from locality (e.g. the Weight-to-Stress Principle (Prince 1992)).

## 7.5 Incremental Learning

The learners presented above were batch learners, and it was shown that for particular input samples, those learners can return the target pattern from limited data in most instances. This section shows how the learners can be modified to become iterative, and shows that they converge to the correct grammar.

I begin this discussion by pointing out one fact about the input samples described earlier. By making the samples consist of words from one to  $n$  syllables, this guarantees in prefix and suffix tree construction that each state fully realises its possible outgoing transitions. I will call a state in a prefix or suffix tree which has every possible incoming and outgoing transition realised (as determined by the sampling language), *saturated*. Because of the kind of sample used to construct the trees, every non-terminal state in the trees in Figure 4 and Figure 6 is saturated.

When we consider iterative learners, it is not the case that every non-terminal state in the prefix and suffix trees will be saturated. For example, if the sample in Table 1 for the LHOR pattern excluded word types  $\acute{H}L$ ,  $\acute{H}LH$ ,  $\acute{H}LHH$ ,  $\acute{H}LHL$ , then the branch in the prefix tree from state 12 to 23 would be missing, and state 12 would be non-final. If  $LL\acute{L}$ ,  $LL\acute{H}$ ,  $LL\acute{H}L$ ,  $LL\acute{H}H$  were also excluded then the branches from states 4 to 5 and 4 to 10 would also be missing. For such a sample, the Forward Learner would merge states 4 and 12 since they would both be non-final, non-start, and each would have only one incoming transition labeled L, and outgoing transition labeled L. Such a merge would be an error since the resulting pattern obtained would include forms like  $\acute{H}LL\acute{H}$ .

---

<sup>27</sup>Though one undesirable consequence of 2-2 neighborhood learning is that patterns describable with feet of size four are predicted to exist.

This suggests that although the learner may obtain the correct grammar when the sample includes all words of length equal to or less than some  $n$ , it does not necessarily succeed for intermediate-sized samples—i.e. those samples which include some proper subset of words of length less than or equal to  $n$ . Equivalently, we might say that the learner’s only chance of success comes when every state in the prefix or suffix tree is saturated. However, it can be shown that if the sample is sufficient (i.e. certain states in the trees are saturated), then same-neighborhood merging of states in the prefix or suffix tree of any larger sample returns an acceptor which recognises a language which is a superset of the target language (see appendix B). Consequently, the following iterative procedure provably converges to the target grammar if it has a sufficient sample.

I illustrate the idea with the Forward Learner. Imagine it obtains one word from the target pattern at each time step, and at each time step the learner is allowed to make a hypothesis. Recall that the learner converges if (1) there is some time step  $t$  where the learner hypothesises a grammar generating the target pattern (2) for all future time steps, the learner’s hypothesis stays the same, and (3)  $t$  can be obtained after finitely many time steps.

We modify the Forward Learner by giving it a memory: essentially, it can maintain the prefix tree in one memory bank, and maintain its most recent hypothesis  $H$  in another. At the next time step, when a word  $w$  is presented, the word is added to the prefix tree. The learner keeps this tree in memory but also merges same-neighborhood states to obtain another acceptor  $H'$ . If  $w$  is not accepted by the most recent hypothesis  $H$ , then  $H$  is discarded and replaced with  $H'$ , and we move on to the next time step. On the other hand, if  $w$  is accepted by  $H$ , then the learner must choose between the current hypothesis  $H'$  and  $H$ . Since  $H$  and  $H'$  are both acceptors, it is possible to check to see whether one pattern is a subset of the other<sup>28</sup>, in which case the acceptor representing the smaller pattern is chosen.<sup>29</sup>

As an illustration, in the case above, suppose that the current hypothesis the learner obtains is based on the sample described, and so incorrectly accepts words like  $\acute{H}LL\acute{H}$ . If the next word heard by the learner is  $\acute{H}LH$ , then the updated prefix tree now distinguishes states 12 from 4, and they will not be merged. Moreover, every word accepted by  $H'$  is also accepted by  $H$ . Thus the learner keeps  $H'$ , which is the more restrictive hypothesis.

It is easy to show that this iterative learner eventually converges to the target patterns the batch learner succeeds on. This is because in those languages, there is a finite sample such that the prefix tree contains non-terminal states which are saturated and that merging same-neighborhood states in this prefix tree returns an acceptor equal to the target pattern.<sup>30</sup> Therefore, by the theorem in Appendix B, merging same-neighborhood-states in the prefix tree built from any larger sample returns a language

---

<sup>28</sup>This can be accomplished by checking whether the intersection of the complement of one acceptor and the other acceptor is empty.

<sup>29</sup>If neither is a subset, it does not matter which is chosen. This case can (provably) only happen finitely many times because there will be some point in the text after which one acceptor will always be a subset of the other.

<sup>30</sup>The reason the learner is guaranteed to see this sample after finitely many time steps is the sample size is finite and each word in the sample is guaranteed to occur at some time step (because the learner is given a positive text from the target language (see Gold (1967))).

which is a superset of the target pattern. Thus the iterative learner above, at this point, rejects any new hypothesis and keeps the hypothesised grammar that generates the target language. Therefore we have exact identification in the limit.

## 8 Comparison to other Learning Models

Here I compare the Forward Backward Learner to the ordered cue-based learner in the Principles and Parameters framework (Dresher & Kaye 1990, Gillis *et al.* 1995), a perceptron-based learner (Gupta & Touretzky 1994), and an OT-based learner (recursive constraint demotion with robust interpretive parsing) (Tesar 1998, Tesar & Smolensky 2000). Like the Forward Backward Learner, each of these learning models was evaluated with respect to stress patterns. However, exact comparisons are not possible because each learner was tested on a different set of stress patterns with different kinds of input samples.

Gillis *et al.* (1995) implement the cue-based model presented in Dresher & Kaye (1990). The ten parameters yield a language space consisting of 216 languages. The language space is based on actual stress patterns but does not include all attested stress types. The learner discovers parameter settings compatible with 75% to 80% of these languages when provided a sample of all possible words from one to four syllables. As Dresher (1999) notes, it is unknown but perfectly conceivable that accuracy increases if longer words are admitted into the sample.

Gupta & Touretzky (1994) present a perceptron with nineteen stress patterns, of which it successfully learns seventeen. The training input consists of a sample of all words of length seven syllables and less, and is presented to the perceptron at least seventeen times. This is the smallest number of times that resulted in successful learning any of the nineteen patterns (e.g. the perceptron learned Latvian, a QI single system with word-initial stress, when presented with such training input). The largest number of presentations of the sample is 255 (for Lakota, a QI single system which places stress on the peninitial syllable). If the perceptron is given a training sample of shorter words, it is able to learn the two patterns which it otherwise fails to learn.

Tesar and Smolensky (2000) report 12 constraints which yields a typology of 124 languages. Like the language space in the P&P model above, this is an artificial language space based on actual languages (i.e. not all attested patterns are included). If the initial state of the learner is monostratal—that is, no a priori ranking—then the learner succeeds on about 60% of the languages. When a particular initial constraint hierarchy is adopted, the learner achieves  $\sim 97\%$  success.

The FBL is certainly simpler than the P&P and OT learners in the sense that it uses fewer a priori parameters. How exactly these a priori parameters are to be counted is not clear since the models are not on a level playing field. But certainly the FBL, which has no a priori P&P parameters or OT constraints, is much simpler. The speed at which the FBL converges (measured by sample size) appears slower than both of these models; this is almost certainly related to the fact that the hypothesis space of the FBL is larger.

When the FBL is compared to the perceptron learner, it is less clear which is the simpler model. However, the perceptron learner is much, much slower than the FBL

as it requires repeated presentations of words.

However, the main advantage the FBL has over the other models is that the locus of explanation of some aspects of the typology of stress patterns now resides in the learning process. In fact (with the one caveat mentioned earlier) we can say that the reason stress patterns are neighborhood-distinct is because learners generalise from their experience in the way predicted by the FBL. In this way, the FBL is more explanatory than the other models, where the locus of explanation lies in the parameters or constraints (which may be derived from other principles or which may be stipulated), or is obfuscated.

## 9 Conclusion

Explaining how children infer grammatical rules based on their limited finite experience is one of the central goals of modern linguistics. Because children and languages are complex and many factors influence acquisition—physiological, sociolinguistic, articulatory, perceptual, phonological, syntactic, semantic—a simpler question is often asked: How could *anything* learn some aspect of language from the kinds of evidence to which children are exposed? In this paper, the aspect under investigation are the stress rules found in the world’s languages. However, the learning problem was factored even further: What contribution can a particular inductive principle—here a particular notion of locality—make to learning stress rules?

To answer this question, two recent surveys (Bailey 1995, Gordon 2002), when put together, yield a typological survey of 423 stress languages and 109 distinct stress patterns. An examination of each stress pattern—represented by a finite state acceptor—revealed that 107 of them are tail or head canonically 1-1 neighborhood-distinct. In other words, the grammars of these stress patterns refer to phonological environments (states) that are uniquely defined by local properties. Furthermore, many logically possible unattested stress patterns are not 1-1 canonically neighborhood-distinct. Thus neighborhood-distinctness approximates the attested typology in a non-trivial way. This leads to one hypothesis put forward in this paper are that all phonotactic patterns (of which stress patterns form a subclass) are canonically 1-1 neighborhood distinct.

Neighborhood-distinctness is not only interesting because it is a novel formulation of locality in phonology and a (near) universal of attested phonotactic patterns, but also because it naturally provides an inductive principle learners can use to generalise. The Forward Backward Neighborhood Learner, which merges same-state neighborhoods in prefix and suffix trees correctly learns 100 of the 109 stress patterns. This learner is interesting for three reasons. First, it is unable to learn many non-neighborhood-distinct patterns, such as logically possible but unattested stress patterns that are describable with feet of size four or more. Indeed, it was discovered that learners which generalise in this way are, in a sense, unable to count past two, thereby deriving the non-counting nature of phonological patterns from this notion of local environment. Secondly, the use of the suffix tree appears particularly suited to stress patterns anchored at the right word edge. Finally, the learner shows that a formulation of locality in phonology makes a significant contribution to learnability as this factor alone is sufficient for identification in the limit from positive data of almost

all stress patterns in the typology. Since investigation of the ‘failure’ cases revealed that the patterns obtained by the learner are either slight overgeneralizations or plausible optional stress patterns involving secondary stress, another hypothesis was put forward: stress patterns fall within the range of the FBL learning function. We conclude that if human learners generalise in the way predicted by the FBL, it can explain certain aspects of the typology of the attested stress patterns.

These results lead to new formal, typological, descriptive and experimental questions for researchers, some of which have already been mentioned:

1. In OT, what are the typological consequences of requiring constraints in CON to be neighborhood-distinct?
2. As we enlarge the stress typology, do additional stress patterns of languages conform to Hypotheses 1 and 2, or not?
3. Are the studies of the attested patterns sufficiently thorough to not allow for other interpretations of the actual data, such as the ones predicted by the FBL?
4. Do adults or children learn neighborhood-distinct patterns more easily than non-neighborhood-distinct patterns?

In addition to these questions, I would like to mention two other avenues of research that appear fruitful. First, where do neighborhood-distinct patterns fall in the sub-regular hierarchy, which categorises regular languages according to their inherent complexity (McNaughton & Papert 1971)? McNaughton & Papert (1971) show that various independent measures of complexity of regular sets—in automata-theoretic terms, in terms of regular expressions, and in terms of logic—show a high degree convergence; that is the same patterns that are complex with respect to one measure are as complex with respect to some other measure. More recently, Pullum & Rogers (2007) argue that the sub-regular hierarchy provides a natural stomping ground to investigate the cognitive abilities of humans and other species. They suggest language-learning experiments be performed (on both children and adults, including those from other species), to see whether the subjects can learn patterns which differ according to those various degrees of complexity. Given that stress patterns (and phonotactic patterns in general) are describable as regular sets, it is reasonable to make them one focus of the inquiry proposed by Pullum and Rogers. Artificial language learning experiments such as poverty of the stimulus experiments (Wilson 2006) appear capable of shedding light on this question.

The other goal is to determine more precisely what kind of sample is needed to guarantee correct generalization by the FBL. It may well be that the sample size needed to identify the pattern exactly in the limit requires more word types than what we may reasonably expect to find in a child’s linguistic environment. If this is indeed the case, it is likely to lead to the discovery of additional factors that plausibly play a role in human language learning.

Gildea & Jurafsky (1996) are an example of the kind of research that this last line of inquiry can lead to. Realizing that the English rule of flapping can be represented with a subsequential transducer—a particular kind of finite-state machine—they asked whether the flapping rule can be discovered from input/output pairs based on words found in the Carnegie Mellon University English Pronouncing Dictionary (CMUDict).

This question is relevant because Oncina *et al.* (1993) show that the rules representable by subsequential transducers are identifiable in the limit from positive data (here, a sequence of input/output pairs). On the other hand, Gildea and Jurafsky show that the flapping rule is not inferred from the data given 50,000 input/output pairs based on CMUDict.<sup>31</sup> They go on to add phonologically-motivated inductive principles to the algorithm given by Oncina *et al.* and show the rule obtained from the same input sample is much closer to the target flapping rule.<sup>32</sup> This example reinforces the point that there are multiple factors in the learning process and that studying the contributions each particular factor makes to learning (e.g. by studying the kinds of input samples each factor requires for correct generalization, or by how combinations of factors reduce the size of the necessary input sample) will lead to new insights.

## A Appendix: Results of the Neighborhood Learning Study

The tables below are interpreted as follows. In the ‘FL’, ‘BL’, and ‘FBL’ columns, circled numbers mean the Forward Learner, the Backward Learner, and Forward Backward Learner identifies the pattern, respectively. The number inside the circle indicates which forms were necessary for convergence. Specifically,  $\textcircled{n}$  means the learner succeeded learning the pattern with a sample of words consisting of every word which obeyed the pattern which was between one and  $n$  syllables in length. The ‘Notes’ column indicates whether or not there are any phonotactic restrictions (which the sample obeys) or other relevant information. In particular, X and Y indicate whether the stress pattern is not tail or head canonically neighborhood-distinct, respectively. Thus, absence of X (Y) indicates tail (head) canonical neighborhood-distinctness. Table 7 provides an explanation of the notes. The ‘Name’ column provides the name of a language in the typology which exemplifies the pattern, which is uniquely identified by the number in the ‘#’ column.

The ‘Main’ column contains the Syllable Priority Code (SPC), which was developed by Bailey (1995) as a shorthand for indicating primary stress assignment rules. The last character of the SPC ( $L$  or  $R$ ) indicates from which edge of the word to begin counting. Thus the initial syllable is designated 1L, the peninitial 2L, the penultimate 2R, and the final syllable 1R. Thus the simplest SPC codes, such as  $1L$  (Afrikaans), simply mean main stress falls on the initial syllable.

Generally, more complex SPCs can be read as a series of if-then-else statements. Slashes indicate a quantity-sensitive rule with rules governing heavier syllables occurring left of the slash. Thus the SPC  $12/2L$  (Maidu) unpacks to the following: If the initial syllable is heavy, it gets stress, else if the peninitial syllable is heavy, it gets

---

<sup>31</sup>These are not conflicting results, it just means the input sample needed for exact identification in the limit is not present in the particular sample used by Gildea and Jurafsky, which arguably offers the learner a richer linguistic environment than what children are exposed to.

<sup>32</sup>Another line of inquiry raised by Gildea and Jurafsky’s work that has not been pursued as far as I know, is investigating what kinds of phonological rules can be learned by their proposals. In other words, what are the typological predictions—i.e. the range—of their learner?

stress, else stress falls on the peninitial syllable. If the numbers are suffixed with @s, it means primary stress is assigned if the syllable position carries secondary stress.

Unbounded patterns, where the stress can fall any distance from the word edge, use the *12..89* construct. For example, the SPC for Amele *12..89/1L* unpacks to the following: If the first syllable counting from the left is heavy then it receives primary stress, else if the second syllable counting from the left is heavy then it receives primary stress . . . otherwise (if there are no heavy syllables) the first syllable counting from the left receives primary stress. Since words are unbounded in length, Bailey (1995) uses *..89* to indicate ‘and so on’ in the increasing order for any length. Thus 89 do not literally mean the 8th or 9th syllable. Rather 9 means the farthest syllable from the relevant edge and 8 means the next-to-farthest syllable from the relevant edge and so on. See Bailey (1995) for more details.

SPCs that are followed by  $(n+)$  means the code only applies to words that have at least  $n$  syllables. Likewise SPCs that are followed by  $(n-)$  means the code only applies to words that have at most  $n$  syllables.

The ‘Secondary’ column contains extensions I made to the SPC in order to describe secondary stress patterns. ‘None’ of course means that no secondary stress is present. ‘Not included’ means that source material reports secondary stresses, but that either 1) the source material did not describe it, usually because it was deemed too complex, or 2) the source material did describe it, but the pattern was either unclear or too complicated for me to incorporate into the study due to the usual suspect: time.

I indicate secondary stress patterns that can be described iteratively with the prefix *i-*. The prefix *i2* means the second syllable from a stress receives a stress (in both directions). The first stress is indicated with a SPC suffixed with a @ symbol. Thus *i2@1L* (Bagandji) indicates secondary stresses fall on odd syllables from the left, whereas *i2@2R* (Anejom) indicates secondary stresses fall on even syllables from the right. @m means that the first stress upon which the iterative procedure is based is the position of main stress. @mL means the iterations proceeds only leftwards of main stress. Likewise, @mR means the iterations proceeds only rightwards of main stress.

When the secondary stress rules are quantity-insensitive, I use H,L,X to designate heavy, light, and either heavy or light syllables, respectively. Thus a typical trochaic pattern is designated  $i('H, 'LL)$  and a typical iambic pattern  $i(H', LX')$ . If the iterative procedure begins from the word edge (as opposed to from a particular position), I forgoe the connective @ and just suffix *L* or *R* to indicate whether the pattern proceeds from the left or right edge, respectively. Thus  $i('H, 'LL)R$  (Inga) means trochees are iteratively constructed from the right word edge.

Whenever only heavy syllables bear secondary stress, I indicate this with *H*. Sometimes it is necessary to explicitly mention that secondary stress only precedes main stress (as in cases describable with foot extrametricality), in which case I use the symbol <.

Table 2: Quantity-Insensitive Single and Dual Patterns

#	Name	Main	Secondary	Note	FL	BL	FBL
SINGLE							
1.	Afrikaans	1L	None		④	④	④
2.	Abun West	1R	None		④	④	④

*Continued on next page*

#	Name	Main	Secondary	Note	FL	BL	FBL
3.	Diegueno (roots)	1R	None	B	④	④	④
4.	Agul North	2L	None		⑤	⑤	⑤
5.	Alawa	2R	None		⑤	⑤	⑤
6.	Mohawk	2R	None	A	⑤	⑤	⑤
7.	Cora	1L (2-), 3R (3+)	None		⑥	⑥	⑥
8.	Paamese	3R (3+), 1L (2-)	None	B,X	×	⑥	⑥
9.	Bhojpuri	3R (4+), 2R (3-)	Not included	X	×	⑥	⑥
10.	Icua Tupi	3R (5+), 2R (4-)	None	X,Y	×	×	×
11.	Bulgarian	lexical	None		④	④	④
DUAL							
12.	Gugu-Yalanji	1L	2R		⑥	⑥	⑥
13.	Sorbian	1L	None (3-), 2R (4+)	X	×	⑥	⑥
14.	Walmatjari	1L	2R or 3R (5+), 2R (4), None (3- )	Y	×	⑥	⑥
15.	Mingrelian	1L	3R (4+), None (3-)	X	×	×	×
16.	Armenian	1R	1L		⑤	⑤	⑤
17.	Udihe	1R	None (2-), 1L (3+)		⑤	⑥	⑥
18.	Anyula	2R	1L (4+), None (3-)		⑥	⑦	⑦
19.	Georgian	3R (3+), 2R (2-)	1L (5+), None (4-)		⑦	⑧	⑧

Table 3: Quantity-Insensitive Binary and Ternary Patterns

#	Name	Main	Secondary	Note	FL	BL	FBL
BINARY							
20.	Bagandji	1L	i2@1L		⑤	⑤	⑤
21.	Maranungku	1L	i2@1L	B	⑤	⑤	⑤
22.	Asmat	1R	i2@1R		⑤	⑤	⑤
23.	Araucanian	2L	i2@2L		⑥	⑥	⑥
24.	Anejom	2R	i2@2R		⑥	⑥	⑥
25.	Cavinena	2R	i2@2R	A	⑥	⑥	⑥
BINARY WITH LAPSE							
26.	Anguthimri	1L	i2@1L, no 1R		⑥	⑥	⑥
27.	Bidyara Gungabula	1L	i2@1L, no 1R	A	⑥	⑥	⑥
28.	Burum	1L	i2@1L, optional no 1R		⑤	⑤	⑤
29.	Garawa	1L	i2@2R, 1L, no 2L		⑥	⑥	⑥

*Continued on next page*

#	Name	Main	Secondary	Note	FL	BL	FBL
30.	Indonesian	2R	i2@2R, 1L, no 2L (4+), None (3-)	X	×	⑧	⑧
31.	Piro	2R	i2@1L, 2R, no 3R		⑥	⑦	⑦
32.	Malakmalak	12@sL (3+), 1L (3-)	i2@2R (3+), None (3-)		⑥	⑥	⑥
BINARY WITH CLASH							
33.	Gosiute Shoshone	1L	i2@1L, 1R		⑤	⑥	⑥
34.	Tauya	1R	i2@1R, 1L		⑥	⑤	⑥
35.	Southern Paiute	2L (3+), 1L (2-)	i2@2L, 2R, no 1R (3+), None (2-)	B, Y	⑦	×	⑧
36.	Biangai	2R	i2@2R, 1L		⑦	⑥	⑦
37.	Central Alaskan Yupik	1R	i2@2L	B	⑥	⑥	⑥
TERNARY							
38.	Cayuvava	1L (2-), 3R (3+)	None (2-), i3@3R (3+)	A, X	×	⑧	⑨
39.	Ioway-Oto	2L	i3@2L		⑦	⑧	⑧

Table 4: Quantity-Sensitive Bounded Patterns

#	Name	Main	Secondary	Note	FL	BL	FBL
LEFTMOST HEAVY OTHERWISE LEFTMOST							
40.	Murik	12..89/1L	None	C	④	④	④
41.	Lithuanian	12..89/1L	None	D	④	④	④
42.	Amele	12..89/1L	None		④	⑤	⑤
43.	Mongolian Khalkha (per Street)	12..89/1L	H		④	⑤	⑤
44.	Yidin	12..89/1L	i2@m	B	⑤	×	⑤
45.	Kashmiri	12..78/ 12..78/1L	None		×	⑥	⑥
46.	Maori	12..89/ 12..89/1L	Not included		⑤	⑤	⑤
47.	Mongolian Khalkha (per Stuart)	12..89/2L	None		⑤	⑤	⑤
LEFTMOST HEAVY OTHERWISE RIGHTMOST							
48.	Komi	12..89/9L	None		④	④	④
RIGHTMOST HEAVY OTHERWISE LEFTMOST							
49.	Kuuku-Yau	12..89/9R	1L, H		⑤	⑤	⑤
50.	Nubian Don- golese	23..89/9R	H		⑤	⑤	⑤
51.	Mongolian Khalkha (per Bosson)	23..891/9R	H		×	⑤	⑤

*Continued on next page*

#	Name	Main	Secondary	Note	FL	BL	FBL
52.	Buriat	23..891/9R	1L, H		×	⑥	⑥
53.	Arabic Classical	1/23..89/ 9R	None		④	④	④
54.	Cheremis East- ern	23..89/9R	None		⑤	⑤	⑤
55.	Chuvash	12..89/9R	None		④	④	④
RIGHTMOST HEAVY OTHERWISE RIGHTMOST							
56.	Golin	12..89/1R	None		⑤	④	⑤
57.	Cheremis Meadow	1/23..891/ 1R	None		⑤	④	⑤
58.	Mam	12..89/12/ 2R	None	B	⑤	⑤	⑤
59.	Klamath	12..89/23/ 3R	if 3R=SH, 2R=H then 2R		×	×	⑥
60.	Seneca	see note	i2@m < m	E	⑦	⑦	⑦
61.	Cheremis Moun- tain	23..89/2R	None		⑥	⑤	⑥
62.	Hindi (per Jones)	23..891/2R	None		×	⑤	⑥
63.	Sindhi	23..891/2R	H		×	⑤	⑥
64.	Bhojपुरi (per Shukla and Tiware)	23..891/2R	'Hm'H, m'LL, 1L	Y	×	×	⑥
65.	Hindi (per Kelkar)	23..891/ 23..891/2R	H, i('LL)@m < m, m <i(LL')@m	X, Y	×	×	×

Table 5: Quantity-Sensitive Bounded Single, Dual, and Multiple Patterns

#	Name	Main	Secondary	Note	FL	BL	FBL
SINGLE							
66.	Maidu	12/2L	Not included		④	④	④
67.	Hopi	12/2L	None	B	④	④	④
68.	English verbs	12/2R	Not included		④	④	④
69.	Kawaiisu	12/2R	None	B	④	④	④
70.	Shoshone Tump- isa	21/1L	Not included		⑤	⑤	⑤
71.	Javanese	21/1R	None		⑤	⑤	⑤
72.	Manobo Sarangani (per Meiklejohn & Meiklejohn)	21/1R	None	B	⑤	⑤	⑤
73.	Awadhi	21/2R	None	B	⑤	⑤	⑤
74.	Malay (per Lewis)	23/3R (3+), 12/2L (2-)	None		×	⑤	⑤
75.	Latin Classical	23/3R (3+), 1L (2-)	None	B	⑤	⑤	⑤
76.	Hebrew Tiberian	12/21/1R	Not included		④	④	④
77.	English (nouns per Pater)	1@w3/234@sR	i('H,'LL)R		⑤	⑤	⑤

*Continued on next page*

#	Name	Main	Secondary	Note	FL	BL	FBL
78.	Arabic Cairene	1@w3/23@sR	None	B	④	④	④
79.	Arabic Dama- scene	1@w3/23R	None		⑤	⑤	⑤
80.	Arabic Cyre- naican Bedouin	1@w3/23@sR (3+), 12/1R (2-)	i(H',LX')L (invs) (3+), None (2-)	B	×	×	×
81.	Hindi (per Fair- banks)	12/2/34@sR (3+), 1L (2-)	i('H,'LL)R (invs) (3+), None (2-)	X	×	×	×
82.	Piraha	123/123/ 123/123/1R	None	X	×	×	×
DUAL							
83.	Maithili	213/2R	1L	B	⑥	⑥	⑥
MULTIPLE							
84.	Cambodian	1R	H	B, G	⑤	⑤	⑤
85.	Yapese	12/1R	H		④	④	④
86.	Tongan	12/2R	H	B	④	④	④
87.	Miwok Sierra	12/2L	H	B	④	④	④
88.	Gurkhali	12/1L	m < H		④	④	④

Table 6: Quantity-Sensitive Bounded Binary and Ternary Patterns

#	Name	Main	Secondary	Note	FL	BL	FBL
BINARY							
89.	Aranda Western	12/2L (3+), 1L (2-)	i2@m, no 1R (3+), None (2-)		⑥	⑥	⑥
90.	Nyawaygi	12@sL	i('H,'LL)R		⑤	⑤	⑤
91.	Wargamay	12@sL	i('H,'LL)R, no 'HL	B, I	⑥	⑥	⑥
92.	Romansh Berguener	12/2R	i('H,'LL)L		⑥	⑥	⑥
93.	Greek Ancient	12/2R	i('H,'LL)R		⑤	⑤	⑤
94.	Fijian	12/2R	i('H,'LL)R	B	⑤	⑤	⑤
95.	Romanian	12/2R	i2@m		⑤	⑤	⑤
96.	Seminole Creek	12@sR	i(H',LX')L	B	⑤	⑤	⑤
97.	Aklan	21/1R	i('H,'LL)@m < m		⑤	⑤	⑤
98.	Malecite / Pas- samaquoddy	23@sR	i(H',LX')L		⑥	×	⑥
99.	Munsee	23@sR (3+), 12/2L (2-)	i(H',LX')L, no 1R (3+), None (2-)		×	⑥	⑥
100.	Cayuga	23@sR (3+), 1/0L (2-)	i(H',LX')L, no 1R (3+), None (2-)	B	⑥	⑥	⑥
101.	Manam	123/23/3R	i('H,'LL)@m < m		⑤	⑤	⑤

*Continued on next page*

#	Name	Main	Secondary	Note	FL	BL	FBL
102.	Arabic Negev Bedouin	1@w3/23@sR (3+), 12/1R (2-)	i(H',LX')L (invs) (3+), None (2-)		×	×	×
103.	Arabic Bani-Hassan	1@w3/23@w2/2R	i('H,'LL)@m < m		⑤	⑤	⑤
104.	Arabic Pales-tinian	1/2/34@sR (3+), 1@w3/9R (2-)	i('H,'LL)L < m	B	×	×	×
105.	Asheninca	234/324@s/324@sR	i('H,'LL)L < m (w2=H)	B, X	×	×	×
106.	Dutch	1@w4/23@sR	i('H,'LL)R		⑤	⑤	⑤
TERNARY							
107.	Estonian	1L	i('HX,'XLL,'LL)L		⑥	⑥	⑥
108.	Hungarian	1L	i('HX,'XLL,'LL)L, no 1R		⑥	⑥	⑥
109.	Sentani	12/2R	i('HX,'XLX)@mL		⑦	⑥	⑦

Table 7: Notes for Stress Patterns in Tables 2 - 6

ID	Note
A	no monosyllables
B	no light monosyllables
C	At most one heavy per word
D	At least one heavy per word
E	Rightmost even nonfinal syllable which is either heavy or followed by a (nonfinal) heavy. If no such syllables are present, none are stressed.
F	Pretonic heavies count as light
G	Light syllables occur only immediately following heavy syllables
I	Heavy syllables only occur initially
X	Not tail canonically distinct
Y	Not head canonically distinct

## B Appendix: Proving Convergence

This appendix provides a lemma and theorem which are needed to establish convergence of the Forward and Backward learners.

**Notation.**  $\pi$  indicates a partition of some set of states  $Q$ .  $B(q, \pi)$  denotes the block of states containing  $q \in Q$  in some partition  $\pi$ . An acceptor  $A/\pi$  is the acceptor acquired by merging states in  $A$  which are in the same in the same block in partition  $\pi$ .  $\pi_{nd}$  is the partition obtained by placing states with the same neighborhood in the same block.

**Lemma 1** Let  $S$  and  $S'$  be finite samples of  $L$  and denote the states of  $PT(S)$  and  $PT(S')$  with  $Q$  and  $Q'$ , respectively. If

1.  $S \subset S'$
2.  $L = L(PT(S)/\pi)$
3.  $\pi$  is a restriction of  $\pi'$  to  $Q$ .

Then  $L(PT(S)/\pi) \subseteq L(PT(S')/\pi')$ .

**Proof:** Note that since  $S \subset S'$ ,  $Q \subset Q'$ . Consider any

$$w = x_0x_1x_2 \dots x_k \in L(PT(S)/\pi)$$

Then there is a path through  $PT(S)/\pi$ :  $B(\lambda, \pi), B(x_0, \pi), B(x_1, \pi), \dots, B(w, \pi)$ . Note that  $B(\lambda, \pi)$  is the initial state in  $PT(S)/\pi$  and  $B(w, \pi)$  is a final state in  $PT(S)/\pi$  by definition of a prefix tree and state-merging. Since  $\pi$  is a restriction of  $\pi'$  to  $Q$ ,  $B(\lambda, \pi), B(x_0, \pi), B(x_1, \pi), \dots, B(w, \pi)$  must also be a path in  $PT(S')/\pi'$ . Therefore,  $w \in L(PT(S')/\pi')$  and the lemma is proved.  $\square$

**Definition 1** We say a finite sample  $S$  of a target language  $L$  is *saturated* if the non-terminals of the  $PT(S)$  if there is no  $w \in L$  such that  $PT(S \cup \{w\})$  changes the neighborhood of the non-terminals in  $S$ .

**Theorem 1** Let  $L$  be any regular language and let  $S$  and  $S'$  be finite samples of  $L$  such that:

1.  $S \subset S'$
2.  $L = L(PT(S)/\pi_{nd})$
3.  $S$  is saturated.

Then,  $L(PT(S)/\pi_{nd}) \subseteq L(PT(S')/\pi_{nd})$ .

**Proof:** Let  $Q$  and  $Q'$  denote the states of  $PT(S)$  and  $Q'$  the states of  $PT(S')$ . Note that for any  $u, v \in Q$ ,  $u, v \in Q'$ , since  $S \subset S'$ . Since  $S$  is saturated, the neighborhoods of  $u$  and  $v$  are the same in  $PT(S)$  and  $PT(S')$ . Consequently,

$$B(u, \pi_{nd}^S) = B(v, \pi_{nd}^S) \text{ iff } B(u, \pi_{nd}^{S'}) = B(v, \pi_{nd}^{S'})$$

Thus  $\pi_{nd}^S$  is a restriction of  $\pi_{nd}^{S'}$ . Therefore by Lemma 1,  $L(PT(S)/\pi_{nd}) \subseteq L(PT(S')/\pi_{nd})$ .  $\blacksquare$   
 $\square$

With Theorem 1 in place, it is simple to prove that the iterative versions of the Forward and Backward Learners can succeed, as explained in §7.5.

## References

- Abrahamson, A. (1968). Contrastive distribution of phoneme classes in Içuã Tupi. *Anthropological Linguistics* **10**:6. 11–21.
- Albro, Dan (1998). Evaluation, implementation, and extension of primitive Optimality Theory. Master’s thesis, University of California, Los Angeles.
- Albro, Dan (2005). A large-scale, LPM-OT analysis of Malagasy. PhD dissertation, University of California, Los Angeles.
- Angluin, Dana (1980). Inductive inference of formal languages from positive data. *Information Control* **45**. 117–135.
- Angluin, Dana (1982). Inference of reversible languages. *Journal for the Association of Computing Machinery* **29**:3. 741–765.
- Bailey, Todd (1995). Nonmetrical constraints on stress. PhD dissertation, University of Minnesota. Ann Arbor, Michigan. Stress System Database available at <http://www.cf.ac.uk/psych/ssd/index.html>.
- Baković, Eric (2004). Unbounded stress and factorial typology. In John McCarthy (ed.), *Optimality Theory in Phonology: A Reader*. Blackwell, London. ROA-244, Rutgers Optimality Archive, <http://roa.rutgers.edu/>.
- Barker, Muhammad (1963). Klamath dictionary, volume 31 of *University of California Publications in Linguistics*. University of California Press, Berkeley.
- Barker, Muhammad (1964). Klamath grammar, volume 32 of *University of California Publications in Linguistics*. University of California Press, Berkeley.
- Bhatt, R. (1989). Syllable weight and metrical structure of Kashmiri. Unpublished Ms., University of Illinois, Urbana.
- Biermann, A. W. & J. A. Feldman (1972). On the synthesis of finite state machines from samples of their behavior. *IEEE Transactions on Computers* **C-21**:6. 592–297.
- Blumer, Anselm, A. Ehrenfeucht, David Haussler & Manfred K. Warmuth (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* **36**:4. 929–965.
- Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* **21**. University of Amsterdam.
- Boersma, Paul & Bruce Hayes (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* **32**. 45–86.
- Brame, Michael (1974). The cycle in phonology: stress in Palestinian, Maltese and Spanish. *Linguistic Inquiry* **5**. 39–60.

- Chafe, Wallace (1977). Accent and related phenomena in the five nations Iroquois languages. In Larry Hyman (ed.), *Studies in Stress and Accent*, volume 4 of *Southern California Occasional Papers in Linguistics*. Department of Linguistics, University of Southern California, 169–181.
- Chomsky, Noam (1957). Syntactic structures. Mouton & Co., Printers, The Hague.
- Chomsky, Noam & Morris Halle (1968). The sound pattern of English. Harper & Row.
- Clark, Robin (1992). The selection of syntactic knowledge. *Language Acquisition* **2**. 83–149.
- Crowhurst, Megan & Lev Michael (2005). Iterative footing and prominence-driven stress in Nanti (Kampa). *Language* **81**:1. 47–95.
- Dresher, Elan (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* **30**. 27–67.
- Dresher, Elan & Jonathan Kaye (1990). A computational learning model for metrical phonology. *Cognition* **34**. 137–195.
- Eisner, Jason (1997a). Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual ACL and 8th EACL*. Madrid, 313–320.
- Eisner, Jason (1997b). What constraints should OT allow? Talk handout, Linguistic Society of America, Chicago.
- Eisner, Jason (1998). FOOTFORM decomposed: Using primitive constraints in OT. In Benjamin Bruening (ed.), *Proceedings of SCIL VIII*, number 31 in MIT Working Papers in Linguistics. 115–143.
- Ellison, Mark (1994a). Phonological derivation in Optimality Theory. In *COLING 94*, volume 2. 1007–1013. Kyoto, Japan.
- Ellison, T.M. (1994b). The iterative learning of phonological constraints. *Computational Linguistics* **20**:3.
- England, Nora (1983). A grammar of Mam, a Mayan language. University of Texas Press, Austin.
- Everett, Daniel (1988). On metrical constituent structure in Pirahã phonology. *Natural Language and Linguistic Theory* .
- Fairbanks, Constance (1981). The development of Hindi oral narrative meter. PhD dissertation, University of Wisconsin, Madison.
- Finley, Sara & William Badecker (2007). Towards a substantively biased theory of learning. In *Proceedings of Berkeley Linguistics Society 33*.
- Frank, Robert & Giorgio Satta (1998). Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics* **24**:2. 307–315.

- Gibson, Edward & Kenneth Wexler (1994). Triggers. *Linguistic Inquiry* **25**:3. 407–454.
- Gildea, Daniel & Daniel Jurafsky (1996). Learning bias and phonological-rule induction. *Computational Linguistics* **24**:4.
- Gillis, Steven, Gert Durieux & Walter Daelemans (1995). A computational model of P&P: Drescher & Kaye (1990) revisited. In Frank Wijnen & Maaike Verrips (eds.), *Approaches to parameter setting*. Vakgroep Algemene Taalwetenschap, Universiteit van Amsterdam, 135–173.
- Goedemans, R.W.N., H.G. van der Hulst & E.A.M. Visch (1996). Stress patterns of the world part 1: Background. HIL Publications II. Holland Academic Graphics. The Hague.
- Gold, E.M. (1967). Language identification in the limit. *Information and Control* **10**. 447–474.
- Goldsmith, John (1994). A dynamic computational theory of accent systems. In Jennifer Cole & Charles Kisseberth (eds.), *Perspectives in Phonology*. Stanford: Center for the Study of Language and Information, 1–28.
- Goldwater, Sharon (2006). Non parametric bayesian models of language acquisition. PhD dissertation, Brown University.
- Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. 111–120.
- Gordon, Matthew (2002). A factorial typology of quantity-insensitive stress. *Natural Language and Linguistic Theory* **20**:3. 491–552. Additional appendices available at <http://www.linguistics.ucsb.edu/faculty/gordon/pubs.html>.
- Gordon, Matthew (2006). Syllable weight: Phonetics, phonology, typology. Routledge.
- Gupta, Prahlad & David Touretzky (1991). What a perceptron reveals about metrical phonology. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. 334–339.
- Gupta, Prahlad & David Touretzky (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science* **18**:1. 1–50.
- Halle, Morris (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In *Linguistic Theory and Psychological Reality*. The MIT Press.
- Halle, Morris & Jean-Roger Vergnaud (1987). An essay on stress. The MIT Press.
- Hammond, Michael (1986). The obligatory branching parameter in metrical theory. *Natural Language and Linguistic Theory* **4**. 185–228.

- Hayes, Bruce (1981). A metrical theory of stress rules. PhD dissertation, Massachusetts Institute of Technology. Revised version distributed by Indiana University Linguistics Club, Bloomington, and published by Garland Press, New York 1985.
- Hayes, Bruce (1995). *Metrical stress theory*. Chicago University Press.
- Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* .
- Heinz, Jeffrey (2006). Learning quantity insensitive stress systems via local inference. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group in Computational Phonology at HLT-NAACL*. 21–30. New York City, USA.
- Heinz, Jeffrey (2007). The inductive learning of phonotactic patterns. PhD dissertation, University of California, Los Angeles.
- Hopcroft, John, Rajeev Motwani & Jeffrey Ullman (2001). *Introduction to automata theory, languages, and computation*. Addison-Wesley.
- Hudson, Joyce (1978). *The core of Walmatjari grammar*. Canberra: Australian Institute of Aboriginal Studies.
- Hyde, Brett (2002). A restrictive theory of metrical stress. *Phonology* **19**. 313–319.
- Hyman, Larry (1977). On the nature of linguistic stress. In Larry Hyman (ed.), *Studies in stress and accent: Southern California Occasional Papers in Linguistics 4*. Dept. of Linguistics, University of Southern California.
- Idsardi, William (1992). *The computation of prosody*. PhD dissertation, MIT.
- Idsardi, William (2005). Calculating metrical structure. In C. Cairns & E. Raimy (eds.), *Representations and Architecture in Phonological Theory*. MIT Press. Available at <http://www.ling.udel.edu/idsardi/work/>.
- Jain, Sanjay, Daniel Osherson, James S. Royer & Arun Sharma (1999). *Systems that learn: An introduction to learning theory (learning, development and conceptual change)*. The MIT Press, 2nd edition.
- Johnson, C. Douglas (1972). *Formal aspects of phonological description*. The Hague: Mouton.
- Jurafsky, Daniel & James Martin (2000). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Prentice-Hall.
- Kager, René (1999). *Optimality Theory*. Cambridge University Press.
- Kaplan, Ronald & Martin Kay (1981). Phonological rules and finite state transducers. Paper presented at ACL/LSA Conference, New York.

- Kaplan, Ronald & Martin Kay (1994). Regular models of phonological rule systems. *Computational Linguistics* **20**:3. 331–378.
- Karttunen, Lauri (1998). The proper treatment of Optimality Theory in computational phonology. *Finite-state methods in natural language processing* . 1–12.
- Kelkar, A. R. (1968). Studies in hindi-urdu i: Introduction and word phonology. Deccan College, Poona.
- Kenstowicz, Michael (1983). Parametric variation and accent in the Arabic dialects. In *Proceedings of CLS 19*. 205–213.
- Kenstowicz, Michael (1993). Peak prominence stress systems Optimality Theory. In *Proceedings of the 1st International Conference on Linguistics at Chosun University*. Foreign Culture Research Institute, Chosun University, Korea.
- Kenstowicz, Michael (1994). Phonology in generative grammar. Blackwell Publishers.
- Klimov, G.A. (2001). Megrelskii yazyk. In M.E. Alekseev (ed.), *Yazyki Mira: Kavkazskie Yazyki*. Moscow: Izdatelstvo Academia, 52–58.
- Kracht, Marcus (2003). The mathematics of language. Mouton de Gruyter.
- Martin, Andrew (2007). The evolving lexicon. PhD dissertation, University of California, Los Angeles.
- McCarthy, John (2003). OT constraints are categorical. *Phonology* **20**. 75–138.
- McCarthy, John & Alan Prince (1993). Prosodic morphology i: Constraint interaction and satisfaction. Technical Report 3, Rutgers University Center for Cognitive Science. Available on Rutgers Optimality Archive, ROA#482-1201. <http://roa.rutgers.edu>.
- McNaughton, R. & S. Papert (1971). Counter-free automata. MIT Press.
- Michelson, Karin (1988). A comparative study of Lake-Iroquoian accent. Dordrecht:Kluwer.
- Mitchell, T.F. (1975). Principles of firthian linguistics. Longman, London. Pp. 75-98.
- Moreton, Elliott (2007). Learning bias as a factor in phonological typology. In Charles Chang & Anna Haynie (eds.), *The Proceedings of WCCFL 27*. 1–9.
- Niyogi, Partha (2006). The computational nature of language learning and evolution. The MIT Press.
- Nowak, Martin A., Natalia L. Komarova & Partha Niyogi (2002). Computational and evolutionary aspects of language. *Nature* **417**. 611–617.

- Oncina, José, Pedro García & Enrique Vidal (1993). Learning subsequential transducers for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**. 448–458.
- Osherson, Daniel, Scott Weinstein & Michael Stob (1986). *Systems that learn*. Cambridge, Massachusetts: MIT Press.
- Payne, Judith (1990). Ashéninka stress patterns. *Amazonian Linguistics* . 185–209 University of Texas Press, Austin.
- Piattelli-Palmarini, Massimo (ed.) (1980). *Language and learning: The debate between Jean Piaget and Noam Chomsky*. Harvard University Press.
- Popper, Karl (1959). *The logic of scientific discovery*. Basic Books, Inc. New York.
- Prince, Alan (1983). Relating to the grid. *Linguistic Inquiry* **14**:1.
- Prince, Alan (1992). Quantitative consequences of rhythmic organization. *CLS* **26**. 355–398. Parasession of the Syllable in Phonetics and Phonology.
- Prince, Alan & Paul Smolensky (1993). *Optimality theory: Constraint interaction in generative grammar*. Technical Report 2, Rutgers University Center for Cognitive Science.
- Prince, Alan & Paul Smolensky (2004). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell Publishing.
- Pullum, Geoffrey & James Rogers (2007). Aural pattern recognition experiments and the subregular hierarchy. In Marcus Kracht (ed.), *Proceedings of 10th Mathematics of Language Conference*. 1–7. University of California, Los Angeles.
- Riggle, Jason (2004). *Generation, recognition, and learning in finite state Optimality Theory*. PhD dissertation, University of California, Los Angeles.
- Sharpe, Margaret (1972). *Alawa phonology and grammar*. Canberra: Australian National University.
- Sipser, Michael (1997). *Introduction to the theory of computation*. PWS Publishing Company.
- Stabler, Edward P. (2007). *Computational models of language universals: Expressiveness, learnability and consequences*. Cornell Symposium on Language Universals.
- Stolcke, Andreas (1994). *Bayesian learning of probabilistic language models*. PhD dissertation, University of California, Berkeley.
- Stowell, T. (1979). Stress patterns of the world, unite! In *MIT Working Papers in Linguistics*. Department of Linguistics, MIT, Cambridge, MA.
- Tenenbaum, Josh (1999). *A bayesian framework for concept learning*. PhD dissertation, MIT.

- Tesar, Bruce (1998). An interactive strategy for language learning. *Lingua* **104**. 131–145.
- Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. MIT Press.
- Tessier, Anne-Michelle (2006). Stages of OT phonological acquisition and error-selective learning. In Donald Baumer, David Montero & Michael Scanlon (eds.), *Proceedings of the 25th West Coast Conference of Formal Linguistics*. Cascadilla Proceedings Project.
- Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM* **27**. 1134–1142.
- Walker, Rachel (2000). Mongolian stress, licensing, and factorial typology. ROA-172, Rutgers Optimality Archive, <http://roa.rutgers.edu/>.
- Wilson, Colin (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* **30**:5. 945–982.
- Yang, Charles (2000). Knowledge and learning in natural language. PhD dissertation, MIT.