Phonology **26** (2009) 303–351. © Cambridge University Press 2009
doi:10.1017/S0952675709990145 Printed in the United Kingdom

*On the role of locality in learning stress patterns**

Jeffrey Heinz
University of Delaware

This paper presents a previously unnoticed universal property of stress patterns in the world's languages: they are, for small neighbourhoods, neighbourhood-distinct. Neighbourhood-distinctness is a locality condition defined in automata-theoretic terms. This universal is established by examining stress patterns contained in two typological studies. Strikingly, many logically possible – but unattested – patterns do not have this property. Not only does neighbourhood-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-neighbourhood environments perceived in the learner's linguistic input – hence learning neighbourhood-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.

1 Introduction

It has been long observed that natural language patterns, despite their extensive variation, are not arbitrary, though stating exact universals has been difficult (Greenberg 1963, 1978, Mairal & Gil 2006, Stabler 2009).

* Much of this research was performed with the support of a UCLA Dissertation Year Fellowship. Revisions to the initial submission were completed with the support of a University of Delaware Research Fund grant. I 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, in 2007 at the University of Delaware, UCLA and Oakland University, and in 2008 at the University of Maryland. 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, Jason Riggle, Sarah VanWagenen, Colin Wilson and the participants of the 2007 fall seminar on learning theory at the University of Delaware. CarolAnn Edie and Greg Busanus provided invaluable assistance in developing the website, and Robert Wilder provided invaluable assistance in double-checking the finite-state representations as well as database management. Finally I would like to thank the four anonymous reviewers as well as the associate editor, Bill Idsardi. Their comments have greatly improved the paper. I apologise if I have forgotten to thank anyone, and accept responsibility for remaining errors.

One reason to be interested in language universals is that they can help solve the language-learning mystery. The idea is simple: universal properties of language guide a learner towards the target language. This idea has its roots in the philosophy of inductive logic, which state that no generalisation from finite data is possible without some inductive principle (Popper 1959; see also Piattelli-Palmarini 1980). This idea brings a subtle but important shift in perspective: the inductive principles of language learners determine the linguistic generalisations and are therefore one of the factors which shape language typology. Once identified, these inductive principles become properties of the learner which *explain* properties of natural language typology. They are what Moreton (2008) calls 'analytic bias'.

This paper presents a previously unnoticed universal property of the stress patterns in the world's languages: they are, for small neighbourhoods, neighbourhood-distinct (these terms are defined in §4 below). Informally, neighbourhood-distinctness is a locality condition on phonological grammars. This universal is established by examining the stress patterns contained in two recent typological studies, Bailey (1995) and Gordon (2002), and is interesting for at least 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 neighbourhood-distinctness unites the attested patterns in a non-trivial way. Third, neighbourhood-distinctness naturally provides an inductive principle which 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. those sets describable with finite-state automata; see also Idsardi 2008), motivates and defines neighbourhood-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) *Hypothesis 1*
All phonotactic patterns are neighbourhood-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, 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.[1]

The second part addresses the significance of this finding, and how it might 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) in which universal phonological constraints lie (cf. Eisner 1997b, McCarthy 2003). This paper however explores another possibility: that stress patterns are neighbourhood-distinct because the learner itself is unable to distinguish between 'same neighbourhood states' present in the learner's linguistic input and thus generalises to neighbourhood-distinct grammars. In this way, the learner explains why attested stress patterns are neighbourhood-distinct.

As it turns out, the proposed learner – called the Forward Backward Neighbourhood Learner – is unable to learn every logically possible neighbourhood-distinct pattern; however, it does succeed on 100 of the 109 patterns in the stress typology. 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) *Hypothesis 2*

Phonotactic patterns are in the range of the Forward Backward Neighbourhood Learner.

It is not known which of Hypotheses 1 and 2 is the stronger (i.e. the 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. The learning study here is best 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.

[1] Heinz (2007) shows other significant classes of phonotactic patterns are also neighbourhood-distinct, including adjacency patterns describable with trigram grammars (also called locally 3-testable languages in the strict sense; McNaughton & Papert 1971) and long-distance phonotactic patterns.

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 2008) in trying to determine to what extent a learners' biases can explain typological facts (see also related discussion in Stabler 2009). Although this work differs from these others in its focus – they investigate how phonetic factors 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 will not only lead to further investigation into the formal properties of these classes of patterns, but will also guide future empirical work – typological, experimental and descriptive.

1.2 Approaches to learning in phonology

This approach to the learning problem – where the generalisation strategy of the learner directly relates to inherent properties of the hypothesis space – is different from the ones taken in the Principles and Parameters (P&P) and optimality-theoretic (OT) frameworks. In those approaches, the proposed learning mechanisms operate over an additional layer of structure provided by the grammatical framework which is independent of any inherent properties of the hypothesis space.

It is easy to see that this is true by recognising that the learning algorithms that have been proposed for P&P and OT grammars are essentially the same no matter which 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 & Smolensky (2000: 5–6) make this 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 formally 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: 64) makes the same point, also with respect to learners proposed in the P&P framework:

> the learning algorithm is independent of the content of the grammar … for example … it makes no difference to the TLA [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.[2]

The approach taken here also differs from the recent work in probabilistic-based learning: approaches based on OT (Boersma 1997, Boersma & Hayes 2001), minimum description length (Ellison 1991, Goldsmith & Riggle, ms), Bayes' Law (Tenenbaum 1999, Goldwater 2006), maximum entropy (Goldwater & Johnson 2003, Hayes & Wilson 2008) 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 generalisation strategy of the learner is what determines the hypothesis space.

The observation above that most learning proposals in phonology do not make use of properties inherent in the hypothesis space is only an observation, not an argument. It is logically possible that human learners only make use of the structure afforded by P&P or OT frameworks, without any attention to the properties of the constraints or parameters which largely determine the shape of hypothesis space. 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 sufficiently explored in models of phonological acquisition.

[2] Dresher (1999: 28) draws a distinction between the Triggering Learning Algorithm 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 Triggering Learning Algorithm in this respect (see Gillis *et al.* 1995 for further discussion on this point).

1.3 Organisation

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 neighbourhood-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 neighbourhood-distinctness should be handled by a theory of phonology. §6 introduces the learning framework, defines the Forward Backward Neighbourhood Learner (FBL) and gives the results of the study. §7 analyses and interprets these results, and §8 shows how the learner can be modified to be made incremental. §9 compares the FBL to other learning algorithms that have been evaluated in the domain of stress. §10 summarises the results and suggests future research directions.

There are two appendices available as supplementary materials to 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.[3]

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 1977a, Halle & Vergnaud 1987, Idsardi 1992, Bailey 1995, Hayes 1995, Gordon 2002, Hyde 2002). Additionally, learning of stress systems has been approached before (e.g. Dresher & Kaye 1990, Goldsmith 1994, Gupta & Touretzky 1994, Gillis *et al.* 1995, Tesar 1998, Tesar & Smolensky 2000), making it possible to compare learners and results.

2.1 Summary of the typology

As was true for earlier researchers, 'stress pattern' refers to the dominant stress pattern of the language, and thus the typology ignores lexical exceptions. The typology also does not distinguish stress patterns if they differ only in the domain of their application – e.g. roots *vs.* phrases or nouns *vs.* verbs. Also, as is fully explained in §3, stress patterns here are conceived as patterns over strings of syllables, not segments.

Combining Bailey's (1995) and Gordon's (2002) stress typologies yields a typology of 422 languages, exhibiting 109 distinct stress patterns, representing over 70 language families.[4]

[3] The appendices are available as supplementary online materials at http://journals.cambridge.org/issue_Phonology/Vol26No02.

[4] 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 languages in Bailey (1995) and Gordon (2002) are included

Stress patterns are broadly categorised into three groups by primary stress placement: quantity-insensitive, quantity-sensitive bounded and quantity-sensitive unbounded. Appendix A organises the stress patterns in the typology into these three categories and shows the extensive variation documented in those studies. For the sake of completeness, I briefly review these categories and some of the variation below, though most of the discussion will undoubtedly be familiar to anyone with anything greater than a passing interest in stress patterns. For further details, the reader is referred to Bailey (1995), Hayes (1995) and Gordon (2002), and the primary source references therein. Kager (2007) provides a thorough overview of the different kinds of patterns.

Quantity-insensitive (QI) stress patterns, extensively reviewed in Gordon (2002), are those in which the statement of the stress rule need not refer to the quantity, or weight, of the syllables. 319 languages in the typology are quantity-insensitive, exhibiting 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, such patterns are called quantity-sensitive.[5] 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.[6] The resulting typology includes 44 patterns which have quantity-sensitive bounded patterns.

Like the QI patterns, QS bounded patterns can be subdivided into single, dual, binary, ternary and 'multiple' types ('multiple' is defined below). 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 in that there is no clear principled upper limit on how

---

in StressTyp, but the latter includes more languages than the ones in the Bailey (1995) and Gordon (2002) combined. The database is available (June 2009) at http://www.unileiden.net/stresstyp. See also Goedemans *et al*. (1996).

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

[6] This was done with assistance from undergraduates Rachel Schwartz and Stephen Tran at UCLA.

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.

On the other hand, 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. It is the non-local aspect of unbounded patterns which make them challenging for learning proposals. 45 languages with QS unbounded stress systems are in the typology, exhibiting 26 distinct patterns.

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 which 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 as far as I know is not the case in Alawa, and so both of these patterns, which differ minimally in this respect, are included in the typology as distinct patterns.

2.3 Unattested stress patterns

Despite the extensive variation recounted above, stress patterns are not arbitrary. There are 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 of four syllables or longer, and on the first (or final) syllable in words of three syllables or fewer. No stress pattern places stress on every fourth or every fifth syllable (cf. the binary and ternary patterns above which place stress on every second or third syllable).[7] 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

[7] Hammond (1987) reports that in Hungarian secondary stress falls on every fourth syllable. However, this claim is controversial (Hayes 1995: 330).

constrained. This is not a new observation – virtually every previous researcher essentially makes the same point. I emphasise it only to spotlight the question: what property or properties do the attested patterns share which separates them from these unattested patterns?

3 Representing phonotactic patterns

3.1 Patterns as sets

A pattern can be represented as a set. All logically possible strings which obey the pattern are in the set, and all the possible strings which do not obey the pattern are not in the set. For example, consider the following two attested stress patterns.

(3) a. Primary stress falls on the initial syllable and there is no secondary stress.
b. Primary stress falls on the final syllable and secondary stress falls on other odd syllables, counting from the right.

Afrikaans (Donaldson 1993) is an example of language said to have the stress pattern of (3a) and Asmat (Voorhoeve 1965) the stress pattern described in (3b). These patterns are given as sets in (4).

(4) a. {*σ́*, *σσ́*, *σσσ́*, *σσσσ́*, *σσσσσ́*, …}
b. {*σ́*, *σσ́*, *σ̀σσ́*, *σσ̀σσ́*, *σ̀σ̀σσ́*, *σσ̀σ̀σσ́*, …}

Following Chomsky (1957), I use the word 'language' interchangeably with 'pattern' to mean sets like those above.

The linguistic competence of speakers of Afrikaans and Asmat with respect to stress patterns can be characterised by knowledge of sets (4a) and (4b) respectively. For example, these speakers can presumably decide whether a logically possible word obeys the relevant pattern or not. For example, a word represented as *σ̀σ̀σ̀σ̀σ̀σ̀σσ́* obeys the stress pattern of (3b), whereas a word with the representation *σσ̀σ̀σ̀σ̀σ̀σσ́* does not. The decision a speaker makes regarding whether a particular word obeys the pattern is akin to deciding whether that word is in set (4b).

One item of interest about the sets in (4) is that they are infinite. It is therefore not possible to write out these sets in their entirety as a list. Grammars, however, are a finite way of writing an infinite set. There are many grammars that can define the sets above. For example, if we use the perfect grid (Prince 1983), have parameters which allows us to align a peak of this grid with the right edge of the word and set the End Rule parameter to Right, we can generate the infinite set in (4b). Different settings of such parameters (along with a parameter which essentially removes secondary stress) generate the stress pattern in (4a). On the other hand, if we adopt

parameters which give us disyllabic iambic feet (*σὸ*) and foot assignment from the right edge, and set a Head Foot parameter to Right, then it is possible to develop a grammar which generates the same infinite set (4b) above (Hayes 1995). Different settings of these parameters again generate (4a). Additionally, a number of studies show how those sets can be generated in OT with different kinds of constraints (Tesar & Smolensky 2000, Gordon 2002, Hyde 2002).

This paper addresses many of the simple, fundamental questions which arise when we view stress patterns – or any linguistic pattern – as sets. What properties do sets such as the ones in (4) have? What properties set possible stress patterns apart from some of the logically possible but phonologically unnatural stress patterns discussed in §2.3? How can such an infinite set (or a grammar which generates it exactly) be acquired from just finitely many examples? In other words, what inductive principles, in the sense of Popper (1959), allow one to obtain set (4a) from the finite set {*σ́*, *σόσ*, *σόσσ*, *σόσσσ*}? These last questions have been studied extensively by the grammatical inference community, whose research the learner presented in §6 draws upon. De la Higuera (2005, in press) provides an excellent introduction to grammatical inference.

3.2 Phonotactic patterns are regular sets

Interestingly, phonotactic patterns like the stress patterns in (4) above are all REGULAR sets. Regular sets are those than can be generated by finite-state acceptors.[8] A finite-state acceptor (FSA) is a grammar which contains a finite number of states, some of which are START states and some of which are FINAL states, along with labelled TRANSITIONS between those states. A FSA is said to GENERATE a string if and only if there is a path from a start state to a final state which 'exhausts' the string.[9] Examples which make this notion clear are given below. There are several excellent introductions to finite-state acceptors, including Partee *et al.* (1990), Sipser (1997), Hopcroft *et al.* (2001) and Coleman (2005), to which I refer readers. The focus on finite-state representations here follows earlier work in phonology (Johnson 1972, Koskenniemi 1983, Ellison 1992, Kaplan & Kay 1994, Eisner 1997b, Frank & Satta 1998, Karttunen 1998, Riggle 2004, Albro 2005).

As examples, consider the FSAs in Fig. 1, which I call FSA Afrikaans (a) and FSA Asmat (b). In the finite-state diagrams in this paper, start

[8] The Chomsky Hierarchy classifies all logically possible patterns (i.e. sets) according to the kinds of grammars that can generate them (Harrison 1978). The major classes in this hierarchy are finite sets, regular sets, context-free sets and context-sensitive sets.

Like the other classes in the Chomsky Hierarchy, there are many ways to define regular sets – as right-branching rewrite grammars, as regular expressions or as those strings which make true statements written in monadic second order logic. See Kracht (2003: ch. 2) for additional characterisations of regular sets.

[9] In addition to 'generate', FSAs are also said to equivalently ACCEPT or RECOGNISE such strings.

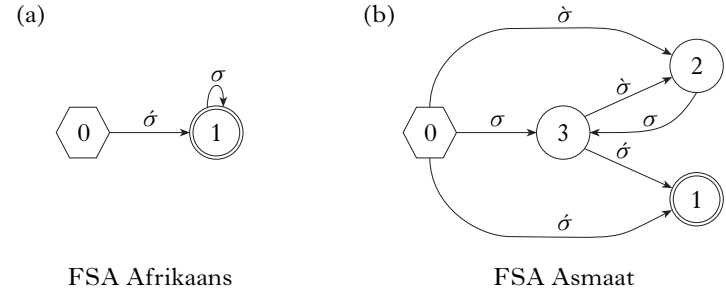


Figure 1
The stress patterns of (a) Afrikaans and (b) Asmat.

states are indicated by hexagons, and final states by circles with double lines, unless mentioned otherwise. The numbers which label the states are irrelevant to the process of generating strings; they are just names of the states to facilitate discussion. FSA Afrikaans generates the string *σ́σσ*. This can be seen by checking whether there is path through the machine that begins at a start state and ends at a final state, along transitions that exhaust *σ́σσ*. FSA Afrikaans is a very simple machine, so it is easy to see that such a path exists. Beginning at state 0, a start state, we travel on the arc labelled *σ́* to state 1. Then we travel along the arc labelled *σ*, which brings us back to state 1. We traverse this arc again returning to state 1 one more time. At this point we have exhausted the string and, since state 1 is a final state, there is in fact a path from a start state to a final state which exhausts the string. FSA Afrikaans does not accept the string *σ́σ́σ*, because there is no start state that has a transition leaving it which is labelled *σ*. It is easy to see that FSA Afrikaans accepts all and only those strings that begin with *σ́* followed by zero or more *σ*. Although FSA Asmat is more complex, in the sense that there are more states and arcs, it is also straightforward to see that the only strings it accepts are the ones which obey the stress pattern of Asmat in (3b).

FSAs Afrikaans and Asmat are faithful representations of the stress patterns in (3), because they generate exactly the sets in (4). Like the generative grammars of earlier researchers, these grammars recognise infinitely many well-formed words. FSAs are categorical phonotactic grammars, because they are devices that can answer yes or no when asked if some logically possible word is possible, which is the minimum requirement of a phonotactic grammar (Chomsky & Halle 1965, 1968, Halle 1978). Stress patterns are usually described as categorical patterns, as opposed to gradient ones, although notice that gradient phonotactic patterns can be described with weighted FSAs.[10]

[10] The FSAs above represent functions which maps any string to yes or no, or equivalently 1 or 0. Hayes & Wilson (2008) provide extensive arguments that many phonotactic patterns are gradient and not categorical. Weighted FSAs, where transitions departing each state are subject to a probability distribution, can be used

In fact, the hypothesis that all phonotactic patterns – not just stress patterns – are regular is well supported.[11] The general form of the argument begins with recognising that phonological alternations can be described as relations. A relation is simply a set of pairs. An underlying form u surfaces as a phonological string s and thus can be written as the pair (u, s). The set of such pairs for a language constitutes the phonological alternations in the language. In the same way that sets can be classified according to their complexity, so can relations. For example, REGULAR RELATIONS are those recognised by finite-state transducers. A finite-state transducer is like an acceptor, except the transitions on the labels are pairs of symbols. Since the righthand side of a regular relation – the set of surface forms s – is a regular set (Sipser 1997, Hopcroft *et al.* 2001), it follows that if all phonological processes which map underlying forms to surface forms can be described with regular relations then all phonotactic patterns are regular sets.

This argument is made by Johnson (1972) and Kaplan & Kay (1994), who show how to construct a finite-state transducer from traditional *SPE*-style ordered rule-based phonological grammars (see also Koskenniemi 1983). Similarly, Gerdemann & van Noord (2000) and Riggle (2004), building on work by Ellison (1994), Eisner (1997b), Frank & Satta (1998), Karttunen (1998) and Albro (2005), show how to construct a finite-state transducer from OT grammars, provided the constraints in Con can be written as vector-weighted finite-state transducers.

For example, consider the (different) OT analyses given in Tesar (1998) and Gordon (2002) of the stress pattern of Asmat (3b). If these analyses were encoded in finite-state OT, applying Gerdemann & van Noord's (2000) or Riggle's (2004) transducer-construction algorithm would yield the (same) acceptor shown above in Fig. 1.[12] In a sense, FSA Asmat (Fig. 1a) implicitly embodies all the constraints and the rankings used in those analyses (and all the parameters and their settings in a P&P analysis, as well as all the rules and orderings in a derivational analysis). This seems to imply that FSAs can serve as a lingua franca between derivational, P&P and OT approaches to phonology.

There are two other reasons to be interested in the hypothesis that all phonotactic patterns are regular. First, the grammatical inference community has developed a considerable literature on learning regular sets. For example, the class of regular languages is not Gold-learnable from positive data (Gold 1967, Angluin 1980), but certain subsets of it are (Angluin 1982, Muggleton 1990, Denis *et al.* 2002, Fernau 2003) (Gold-learnability is explained in more detail in §6). Thus it becomes possible

---

to model gradient phonotactic patterns. In this case, the FSAs are functions mapping strings to real numbers.

[11] This contrasts with syntactic patterns, which are argued to be more complex (Chomsky 1957, Shieber 1985, Kobele 2006).

[12] Since Tesar (1998) uses constraints to place metrical feet in strings, the acceptors above are obtained after removing foot boundaries (which presumably are unpronounced anyway).

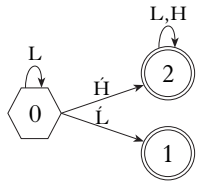


Figure 2
The 'leftmost heavy, otherwise rightmost' stress pattern.

to ask: what subset of the regular languages delimits the class of possible phonotactic patterns and do the properties of this class provide inductive principles for learners? Hypotheses (1) and (2) are essentially one answer to this question.

Second, insights made with respect to regular sets 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 (see also Roark & Sproat 2007 for a finite-state approach to reduplication). Also, if the working assumption that phonotactic constraints are categorical is relaxed, finite-state automata whose transitions are labelled with real-valued weights (i.e. stochastic or probabilistic FSAs) are a natural extension, in which gradient well-formedness patterns can be described.

3.3 QS patterns

The patterns of Afrikaans and Asmat above are QI patterns. QS stress patterns can be represented with FSAs as well. For example, consider the unbounded stress pattern 'leftmost heavy, otherwise rightmost' in (5).

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

(6) shows all strings with four syllables or fewer which obey the pattern in (5). Fig. 2 shows a finite-state acceptor which represents this pattern. Again, it is worthwhile to take a moment to verify that the acceptor in Fig. 2 accepts only those strings which obey the pattern in (5) and rejects all strings which do not obey it.

(6) *Words of four or fewer syllables which obey the 'leftmost heavy, otherwise rightmost' stress pattern*

Ĺ	L	ĹL	ĹH	LĹ
LĹ	ĹLL	ĹLH	ĹHL	ĹHH
LĹL	LĹH	LĹL	LLĹ	LĹLL
LĹLH	ĹLLL	ĹLLH	ĹHLL	ĹHLH
LĹHL	LĹHH	ĹLHL	ĹLHH	ĹHHL
ĹHHH	LLĹL	LLĹH	LLLĹ	LLLĹ

The set of symbols on the transitions in the acceptor in Fig. 2 is not the same as the one in Fig. 1. This set of symbols is referred to as the ALPHABET. There are some important issues regarding the choice of alphabet. To avoid a loss of continuity, this discussion is postponed until §6, where it becomes relevant.

To sum up this section, it has been established that phonotactic patterns such as stress patterns can be represented faithfully with finite-state acceptors.

4 The neighbourhood-distinct hypothesis

In this section, neighbourhood-distinctness is defined and shown to be a locality condition for grammars.

4.1 Locality in phonology

It is generally agreed that locality is an important feature of phonological grammars. McCarthy & 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) refers to 'the well-established generalization that linguistic rules do not count beyond two'.

Within the particular domain of stress, the thinking is no different. Halle & Vergnaud (1987: ix) note that 'it was felt that phonological processes are essentially local and that all cases of nonlocality should derive from universal properties of rule application'. 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. There are two reasons for the attention given to it here: (i) to see in precisely what way the stress patterns in the typology of stress are local, and (ii) to obtain a clear understanding of the contribution this *a priori* notion of locality can make to learning.

4.2 Definition of the neighbourhood

The concept 'neighbourhood' aims to capture the insight that phonological environments are defined locally, while 'distinct' aims to capture the notion that these environments are unique. The main idea is that each state in a finite-state acceptor represents a phonological environment the grammar must be sensitive to. For example in Fig. 1a, state 0 of FSA Afrikaans represents the environment at the beginning of a word, and state

1 represents every other environment (for some related discussion, see Riggle 2004). These two classes of environments are the ones any grammar generating the initial stress pattern of Afrikaans must be sensitive to.

Given the idea that phonological environments are 'local', we identify each state with its local characteristics. Thus the neighbourhood of a state is defined as in (7) (to be revised).

(7) a. the set of incoming symbols to the state
    b. the set of outgoing symbols from the state
    c. whether it is a final state or not
    d. whether it is a start state or not

The neighbourhood of a state can be determined by looking solely at whether or not it is final, whether or not it is a start state, the set of symbols coming into the state and the set of symbols departing the state. Pictorially, all the information about the neighbourhood 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 Fig. 3 belong to some larger acceptor. We can decide that states *p* and *q* have the same neighbourhood because they are both non-final, non-start 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*}.

Figure 3
Two states with the same neighbourhood.

The size of the neighbourhood can be parameterised by adjusting parts (a) and (b) of the definition in (7). 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. I call a path of length *k* a *k*-path. Thus we define the *j*-*k* neighbourhood as in (8).

(8) a. the set of incoming *j*-paths to the state
    b. the set of outgoing *k*-paths from the state
    c. whether it is a final state or not
    d. whether it is a start state or not

It is now possible to define acceptors that are *j*-*k* neighbourhood-distinct.

(9) An acceptor is said to be *j*-*k* NEIGHBOURHOOD-DISTINCT iff no two states have the same *j*-*k* neighbourhood.[13]

The class of neighbourhood-distinct languages is defined in (10).

(10) The *j*-*k* NEIGHBOURHOOD-DISTINCT LANGUAGES are those for which there is an acceptor which is *j*-*k* neighbourhood-distinct.

When the values of *j* and *k* are understood from context, I just write neighbourhood-distinct.

For example, FSA Asmat is 1-1 neighbourhood-distinct. This can be determined by listing the neighbourhood of each state and checking to make sure that each neighbourhood is unique. Table I shows the neighbourhoods of each state in FSA Asmat; no neighbourhood is repeated twice.

state	incoming 1-paths	outgoing 1-paths	final?	start?
0	$\emptyset$	{σ, σ̀, σ́}	*no*	*yes*
1	{σ́}	$\emptyset$	*yes*	*no*
2	{σ̀}	{σ}	*no*	*no*
3	{σ}	{σ̀, σ́}	*no*	*no*

Table I
The 1-1 neighbourhoods of FSA Asmat in Fig. 1b.

Examples of non-neighbourhood-distinct patterns are given by logically possible unattested stress patterns such as those which place stress on every fourth, fifth, sixth or *n*th syllable, which are not 1-1 neighbourhood-distinct. To see why, consider the acceptor in Fig. 4, which generates the logically possible stress pattern which assigns stress to the initial syllable and then secondary stress to every fourth syllable.

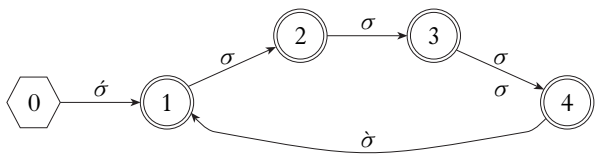


Figure 4
The FSA for a quaternary stress pattern.

[13] 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.

This acceptor generates the strings in (11).

(11) {*σ́*, *σσ́*, *σσσ́*, *σσσσ́*, *σσσσ̀σ́*, *σσσσ̀σσ́*, *σσσσ̀σσσ́*, *σσσσ̀σσσσ́*, …}

In the acceptor in Fig. 4, states 2 and 3 have the same neighbourhood. 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 neighbourhood (because the pattern requires exactly three unstressed syllables between stresses). In contrast, the ternary pattern of Iowoy-Oto, for example, which stresses the peninitial syllable and every third syllable afterwards as shown in (11), is neighbourhood-distinct, as can be easily verified from Fig. 5.

(12) {*σ́*, *σσ́*, *σσσ́*, *σσσσ́*, *σσσσ̀σ́*, …}

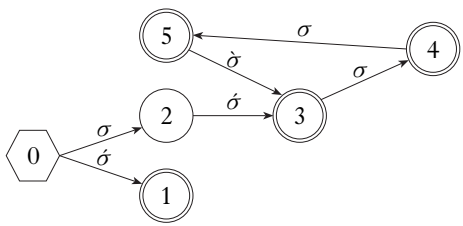


Figure 5
The FSA for the stress pattern of Ioway-Oto.

Generally speaking, neighbourhood-distinctness characterises a locality condition, which is shared by binary and ternary patterns, but not quaternary patterns or higher *n*-ary patterns. In §§4.5–4.6 below, I establish that almost all stress patterns, including the long-distance unbounded types, are neighbourhood-distinct. But first, there is some technical ground to cover.

4.3 Neighbourhood-distinct acceptors and languages

Neighbourhood-distinct languages are defined as those that can be generated by neighbourhood-distinct acceptors. However, many different acceptors can recognise exactly the same language, and it is useful to distinguish certain types of acceptors.

For example, Fig. 6 shows another acceptor which represents exactly the stress pattern of Asmat (3b) (the reader can verify that this is true). Like FSA Asmat (Fig. 1), the acceptor in Fig. 6 has a path from a start state to a final state which exhausts the string for any string which obeys the Asmat pattern. What makes this acceptor unlike FSA Asmat is that there are two start states, and so the acceptor is said to be nondeterministic, because when determining whether the acceptor recognises a particular string there is a choice of which state to begin with.

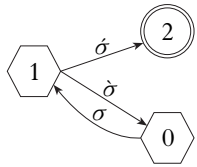


Figure 6
A head canonical acceptor for the stress pattern of Asmat.

Although there are many acceptors which generate a particular set, some are more useful than others. An acceptor like FSA Asmat in Fig. 1b is FORWARD DETERMINISTIC. This means it has one start state, and for each state q in the machine and each symbol a in the alphabet, there is at most one transition labelled a departing q. A forward deterministic acceptor with the fewest states for a language is called the language's TAIL CANONICAL ACCEPTOR, and typically regular patterns are represented with this acceptor. However, another algebraically equivalent choice is the HEAD CANONICAL ACCEPTOR (Heinz 2007). The head canonical acceptor is the smallest REVERSE DETERMINISTIC acceptor recognising some pattern. An acceptor is reverse deterministic if and only if there is at most one final state, and for every state q in the acceptor and each symbol a in the alphabet, there is at most one transition labelled a entering q. The acceptor in Fig. 6 is a head canonical acceptor for the pattern in (3b) (the reader may verify that it is reverse deterministic). The tail and head canonical acceptors can be computed from any acceptor which recognises a given pattern (Hopcroft *et al.* 2001).

The head canonical acceptor for the Asmat pattern is more compact than the tail canonical acceptor, as it has one fewer state and fewer transitions. This is probably due to the fact that it is a right-edge based pattern. Generally, right-edge based stress patterns have smaller head canonical acceptors, whereas left-edge based stress patterns have smaller tail canonical acceptors. This is an interesting observation, which becomes relevant in §8.

I refer to tail canonical acceptors which are j-k neighbourhood-distinct (and the languages they recognise) as TAIL-CANONICALLY j-k NEIGHBOURHOOD-DISTINCT. It is useful to refer to head canonical acceptors which are neighbourhood-distinct (and the languages they recognise) as j-k HEAD-CANONICALLY NEIGHBOURHOOD-DISTINCT. Finally, I refer to patterns which are either tail or head canonically neighbourhood distinct simply as j-k CANONICALLY NEIGHBOURHOOD-DISTINCT. These notions will become useful in §4.5.

4.4 Properties of neighbourhood-distinct languages

An analysis of neighbourhood-distinct languages based on its component parts is begun in Heinz (2007, 2008). The results so far show that the class

is finite, that the *j*-*k* neighbourhood-distinct languages form a structured hierarchy of classes and that it does not have certain closure properties. These are discussed in turn.

1-1 neighbourhood-distinctness is restrictive. The neighbourhood-distinct languages not only form a proper subset of the regular languages over some alphabet Σ, there are only a finite number of them: all regular languages whose smallest acceptors have more than $2^{2|\Sigma|+1}$ states cannot be 1-1 neighbourhood-distinct (since at least two states would have the same neighbourhood). Thus most regular languages are *not* 1-1 neighbourhood-distinct.[14]

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

Finally, Heinz (2007: ch. 6) shows that the *j*-*k* neighbourhood-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.[15]

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

[14] By similar reasoning, one can see that for any particular values of *j* and *k*, most regular patterns are not *j*-*k* neighbourhood-distinct.

[15] These machines are available (June 2009) as part of a stress typology database available at the author's website (http://phonology.cogsci.udel.edu/dbs/stress).

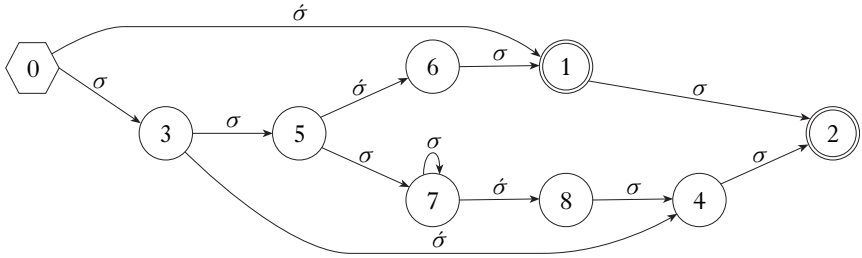


Figure 7
The stress pattern of Içuã Tupi.

To sum up, with two controversial (see below) exceptions, 1-1 neighbourhood-distinctness is a universal property of attested stress patterns. It unites the attested single, dual, binary, ternary and unbounded properties, to the exclusion of quaternary and higher *n*-ary patterns.

4.6 Discussion

How seriously do the two languages which are not canonically 1-1 neighbourhood-distinct challenge the hypothesis that all phonotactic patterns are canonically 1-1 neighbourhood-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 of 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 fewer 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. Figure 7 shows the tail canonical FSA for this stress pattern (states 6 and 8 have the same neighbourhood).

There are other possibilities which render the Içuã Tupi pattern canonically 1-1 neighbourhood-distinct. One possibility is that stress may optionally be placed on the penult or the antepenult in words with five or more syllables (as in Walmatjari; Hudson 1978). Although it may be unfair to assume this (as we can expect Abrahamson to have noted it), this alteration makes the pattern neighbourhood-distinct. In Fig. 7, this change would amount to eliminating states 7 and 8 and their associated transitions, and adding a transition from state 5 to itself, labelled σ.

In the case of Kelkar's (1968) description of Hindi, it is important to recall that the stress pattern of Hindi has been the subject of many different proposals, and there is little consensus as to what the pattern actually is (Ohala 1977, Hayes 1995). Hayes (1995: 162–163, 178) points out that Fairbanks' (1981) analysis is based on additional evidence (metrics), whereas Kelkar's evidence relies on subjective intuitions, which differ from those published in Sharma (1969) and Jones (1971). (The patterns described by Fairbanks, Jones and Sharma are all 1-1 neighbourhood-distinct.)

Nonetheless, even if each different description of Hindi were correct (perhaps because speakers belong to different dialect groups), as in the Içuã Tupi case, a small change to Kelkar's (1968) description renders it 1-1 neighbourhood-distinct. According to Kelkar, Hindi is a QS unbounded system with a three-way quantity distinction, with primary stress falling on the rightmost (non-final) superheavy, or if there are none, on the rightmost (non-final) heavy syllable, or in words with all light syllables, the penult. Secondary stresses fall on heavy syllables, and on alternate light syllables on both sides of the primary stress. This may also be the most complicated pattern in the typology, as measured by the number of states in its tail and head canonical acceptors: 32 and 29 respectively (cf. Pirahã, which has 33 and 18). 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. The size of the relevant FSAs prohibits inclusion here, but 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 neighbourhood-distinct.

To sum up, it is premature to reject the hypothesis that all patterns are canonically 1-1 neighbourhood-distinct because of the counterexamples of Içuã Tupi and Hindi (in Kelkar's analysis). 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 neighbourhood-distinct. Finally, the hypothesis that all stress patterns are canonically 1-1 neighbourhood-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 neighbourhood-distinctness? The issue is raised because neighbourhood-distinctness is a constraint on the well-formedness of an aspect of the total grammar – here, the stress domain. One logical possibility is that neighbourhood-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 neighbourhood-distinct.

In OT, one approach might be to require that individual constraints be neighbourhood-distinct (cf. Eisner 1997b, McCarthy 2003). Recall that under finite-state implementations of OT, a constraint is represented as a finite-state transducer (Eisner 1997a, Frank & Satta 1998, Riggle 2004). Thus this approach requires that the notion of neighbourhood-distinctness be translated from acceptors to transducers, something which can be done reasonably, though certain details will have to be worked out.[16]

This proposal is interesting and, if pursued further, the following remarks apply. First, most phonological constraints are neighbourhood-distinct. The most obvious constraints which are not neighbourhood-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 neighbourhood-distinct constraints would have the desirable effect that many of the unattested patterns – like those describable with feet with four or more syllables – 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 neighbourhood-distinct languages, simply ensuring the individual constraints are neighbourhood-distinct does not guarantee the neighbourhood-distinctness of the stress domain as observed above. It is a worthwhile endeavour to determine (i) whether concrete proposals of OT constraints from such a restricted Con predict a typology where every pattern is neighbourhood-distinct, (ii) what properties are present which ensure this outcome, (iii) if not, whether the Içuã Tupi and Hindi patterns can be explained in this way, and so on. To sum up, it appears that restricting Con to include only neighbourhood-distinct constraints can help explain the neighbourhood-distinctness of stress patterns, but there are plenty of unanswered questions, and pursuing them may prove fruitful.

This paper pursues another explanation of the neighbourhood-distinctness of stress patterns. Namely, stress patterns are neighbourhood-distinct because the learner itself is unable to distinguish between the same-neighbourhood 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.

[16] 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 neighbourhood.

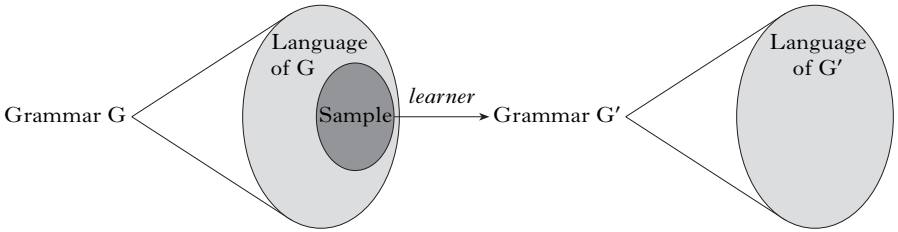


Figure 8
The learning framework.

6 Learning neighbourhood-distinct patterns

6.1 The learning framework

It is useful to make the learning framework, schematised in Fig. 8, explicit. The idea is that the target language is generated from some grammar G. The learner, however, does not hear every word 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' = Language of G? The paradigm schematised in Fig. 8 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. Nowak *et al.* (2002) provide an accessible overview of this learning framework and de la Higuera (in press) provides an excellent in-depth introduction to this learning framework and others (see Jain *et al.* 1999 and Niyogi 2006 for more detailed presentations).

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, Angluin 1980). In particular, no learner can identify the class of all regular languages in the limit from positive data.[17] This result holds even though learners (i) are given only clean, non-noisy learning data and (ii) may take up as much time or space as they like before returning a hypothesised grammar. The utility of a framework that makes these choices is it focuses the problem squarely on generalisation and whether proposed inductive principles allow one to generalise in the desired ways at all. The fact that, despite the generous conditions learners

[17] This result holds even in other learning frameworks with different criteria for success, e.g. the Probably Approximately Correct framework (Valiant 1984; see Anthony & Biggs 1992 and Kearns & Vazirani 1994 for overviews of this framework).

are allowed to operate under, no learner can learn an unrestricted hypothesis space from positive data makes us confront the problem of generalisation directly.

It follows that for learning to be able to take place at all, G' is not drawn from an unrestricted set of possible grammars. The hypotheses available to the learner ultimately determine the kinds of generalisations made and the range of possible natural language patterns. Under this perspective, UG is this set of available hypotheses.

Given that neighbourhood-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 neighbourhood-distinct languages (see §4.4), there are actually very many learning functions which can identify this class of patterns in the limit (Jain *et al*. 1999). 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 Neighbourhood Learner presented below is such a learner, because it generalises to neighbourhood-distinct patterns by its inability to distinguish same-neighbourhood states. Note that to the extent this learner succeeds, it explains why stress patterns are neighbourhood-distinct.

6.2 Overview of the proposed learner

Here I introduce a simple learner which only uses the concept of neighbourhood to generalise. The idea is to merge same-neighbourhood states in the finite-state representation of the input (cf. Angluin 1982, Oncina *et al*. 1993). In particular the algorithm is one instantiation of a general algorithm given by Muggleton (1990: ch. 6), who provides an accessible introduction to algorithms of this sort (see also de la Higuera, in press: chs 4, 11). As it turns out, this learner does not identify the class of neighbourhood-distinct patterns in the limit, though it does succeed for most of the attested patterns. Code to run the learner is available from the author's website (see note 15).

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

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, Muggleton 1990). 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.

For example, consider Fig. 9, which shows two prefix trees. The tree in Fig. 9a is a finite-state representation of linguistic experience consisting of all words up to length three syllables. If we add the word *σσσσ* to this finite sample, we obtain the tree in Fig. 9b.

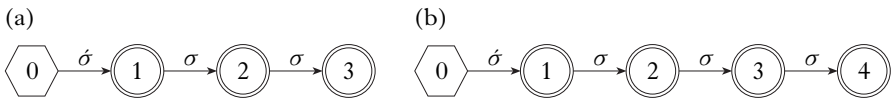


Figure 9

Prefix trees generated by words from the initial stress pattern.
(a) PT accepting {*σ́*, *σ́σ*, *σ́σσ*; (b) PT accepting {*σ́*, *σ́σ*, *σ́σσ*, *σ́σσσ*}.

Figure 10 shows the prefix tree constructed from words up to eight syllables in length from the Asmat pattern.

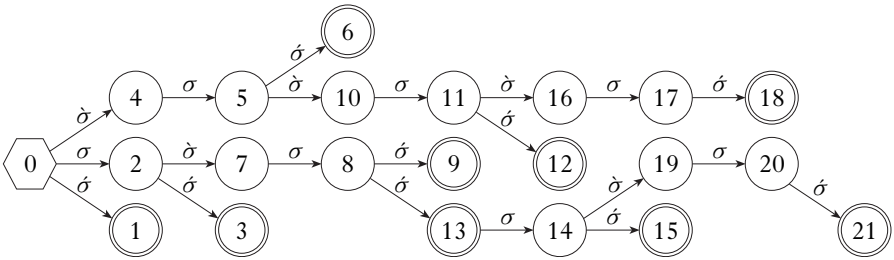


Figure 10

Prefix tree for Asmat words of eight syllables or fewer.

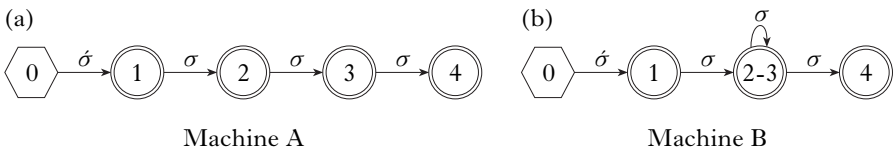
The prefix tree accepts only the finitely many forms that have been observed. There has been no generalisation at this point. These simple examples show that there is structure in the prefix trees and this structure repeats itself. State-merging, described below, can eliminate this redundant structure, possibly leading to generalisation.

I denote the function which maps some finite sample S to a prefix tree PT, which accepts exactly S with PT. Note that $PT(S)$ can be computed efficiently in the size of the sample S (Angluin 1982, Muggleton 1990).

6.4 State-merging as a generalisation 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). One key concept behind state-merging is that transitions are preserved. Another is that if a final (or start) state is merged with a non-final (or non-start) state, then the merged state is final (or start) (Angluin 1982, Muggleton 1990, Hopcroft *et al*. 2001). Generalisations may therefore occur as a consequence of state-merging – because the post-merged machine accepts everything the pre-merged machine accepts, and possibly more.

For example, in Fig. 11 Machine B is the machine obtained by merging states 2 and 3 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 2 to state 3 in Machine B. There is such a transition, but because states 2 and 3 are the same state the transition is now a loop. Whereas Machine A only accepts only four words {*ός*, *όσ*, *όσσ*, *όσσσ*}, Machine B accepts an infinite number of words {*ός*, *όσ*, *όσσ*, *όσσσ*, *όσσσσ*, …}.



Machine A Machine B

Figure 11
An example of generalisation by state-merging.
(a) Machine A; (b) Machine B.

The machines in Fig. 11 are familiar. Machine A is the prefix tree in Fig. 9b and Machine B generates the same language as FSA Afrikaans in Fig. 1a. Note that states 2 and 3 are the only states with the same neighbourhood in Machine A, and these are the states that are merged. This is the central insight that the learner below makes use of: merging same-neighbourhood states in a structured representation of the input can result in a correct generalisation.

The merging process itself 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. If other states had been merged in Machine A in Fig. 11, e.g. states 0 and 4, a different (in this case, incorrect) generalisation would have been made. If all states are merged, the result is a one-state machine, with all transitions self-looping to this state. This machine accepts any string of symbols, indicating a massive overgeneralisation. Thus, choosing which states are to be merged determines the kinds of generalisations that occur (Muggleton 1990,

Heinz 2008).[18] A merging strategy is thus a generalisation strategy. It is an inductive principle, in the sense of Popper (1959).

There is one key result regarding state-merging: 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 acceptor A (Angluin 1982). 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) can be learned with a state-merging strategy.

6.5 The Forward Neighbourhood Learner

The Forward Neighbourhood Learner (FL) merges states in the prefix tree which have the same 1-1 neighbourhood. I use M_{nd} to denote the function which maps an acceptor A to the neighbourhood-distinct acceptor obtained by merging all states in A with the same 1-1 neighbourhood. Note that computing M_{nd} is efficient in the size of A. This is because (i) merging two states is efficient (Hopcroft *et al*. 2001), (ii) at most all pairs of distinct states need be checked for neighbourhood-equivalence to determine if they should be merged and (iii) determining the neighbourhood-equivalence of two states is efficient. It is now possible to state the Forward Neighbourhood Learner precisely.

(13) *Algorithm 1: The Forward Neighbourhood Learner*
Input: a positive sample S
Output: an acceptor A
Let $A = M_{nd}(PT(S))$ and output acceptor A

Figure 11 showed one example of how merging same-neighbourhood states in a prefix tree can lead to successful generalisation. Adding additional words to the prefix tree does not change this result, as the additional states add no new neighbourhoods. Figure 12 shows the result of applying the Forward Learner to the Asmat prefix tree in Fig. 10.

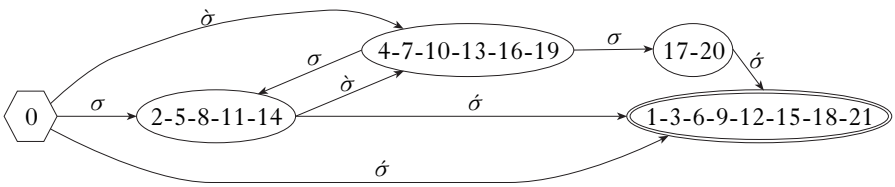


Figure 12
The acceptor obtained by merging same-neighbourhood states in Fig. 10.

[18] In fact, *n*-gram based learning (see e.g. Jurafsky & Martin 2000) can be described exactly as a particular state-merging procedure (Garcia *et al*. 1990, Heinz 2007).

This machine accepts the same language as FSA Asmat, though it is not exactly the same machine. In general, the acceptors obtained by the Forward Learner are not necessarily the same as the tail canonical acceptors, nor are they even deterministic. Machine B in Fig. 11 has a source of non-determinism at state 2–3, and the machine in Fig. 12 has a source at 4-7-10-13-19.

However what matters is whether the *languages* generated by the acceptors are the same as the target patterns. In both examples above, the obtained acceptors do generate the target languages exactly. This can be easily seen by recognising that it is possible to transform these acceptors to the tail canonical ones without adding or losing words to the languages they recognise. In Machine B, for example, state 4 (and its associated transition) can be removed. In Fig. 12, state 17-20 can be removed, along with the transitions associated with it.

6.6 Evaluating the learners

Because the learners in this paper are described in automata-theoretic terms, it is possible that future work will provide proven theorems characterising exactly their behaviour (see Heinz 2008 for the beginning of such an analysis). But since those characterisations are presently unknown, the proposed learners, such as the Forward Learner above, are evaluated in simulations in the following manner. The input samples presented to the learner consist of all words from one to n syllables which obey the stress pattern.[19] 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, this was considered a failure. In such cases, 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 show if the learner succeeded for some input sample of size n, then it also succeeded for sample of size m, where $9 > m > n$. 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. For example, Table III in Appendix A (see note 3) shows that the Forward Backward Neighbourhood Learner succeeds in learning the stress pattern of Amele when provided a sample which consisted of all possible words with one to five syllables. Because this language distinguishes light from heavy syllables, there are $2^1 + 2^2 + 2^3 + 2^4 + 2^5 = 62$ word types in the sample in the simulation.

[19] This decision was made primarily for convenience. The alternative is to consider every subset of words from one to n syllables – an impractical task.

Although this suggests that the learner's success depends on 'knowing' the weight distinction beforehand, this is not so. The choice of alphabet makes the simulation faster to run on a computer, but does not change the nature of the problem. To see why, consider the initial stress pattern in Fig. 1, where the alphabet consists of {σ́, σ}. Suppose we changed the alphabet to include the following symbols {V, CV, CVC, CVV, V́, CV́, CV́C, CV́V}. The machine would now look like the one in Fig. 13.

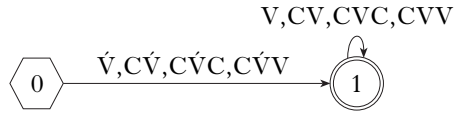


Figure 13
The initial stress pattern with multiple syllable types.

The first thing to notice is that changing the alphabet like this does not change the character of the neighbourhoods in the target grammars. This is important, because the learners operate only on the basis of what distinguishes neighbourhoods. Therefore, with such an alphabet, the learners in fact return a machine like the one in Fig. 13. Thus it is fair to say that the learners discover that the distinctions between the syllable types are irrelevant.

It is true that as the alphabet gets larger, the size of the sample also becomes larger. For example, if we use the four-way syllable distinction above, there would be $4^1 + 4^2 + 4^3 + 4^4 + 4^5 = 1364$ word types in a sample consisting of all words from one to five syllables. Given that what counts as a heavy or light syllable varies across languages, an alphabet which includes every distinction made in any language may get quite large. As the syllable inventory gets larger, so does the concern whether the size of the sample is reasonable, i.e. likely to be present in the lexicon available to a child. This issue is a matter of future research, since it is separate from whether the learners here can be said to discover whether the target pattern is QI or QS when given an alphabet which does not make the distinction overtly.

With these evaluation procedures explained, it can be stated that the Forward Learner successfully learns 85 of the 109 pattern types, as shown in Appendix A.

6.7 Interpreting the results of the Forward Learner

These results also show that the languages in the range of the learning function are not the same as the neighbourhood-distinct languages. The two classes of languages clearly overlap, but the Forward Learner does not identify the class of neighbourhood-distinct languages in the limit. The Forward Learner does not even identify the tail-canonically neighbourhood-distinct languages, falsifying the conjecture of 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.[20]

When we examine the languages which the Forward Learner fails to learn we find that the error is always one of overgeneralisation. 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.

For example, the dual stress pattern of Lower Sorbian requires words with four or more syllables to place primary stress initially, and secondary stress on the penult (e.g. *σσσσ*). However, the Forward Learner returns an acceptor which can stress such words in one of two ways: either the same as the target pattern or by placing primary stress on the initial syllable and having no secondary stress at all (e.g. both *σσσσ* and *σσσσ* are generated). The reader is referred to Heinz (2006) for details.

Another characteristic that all stress patterns for which the Forward Learner fails (except Kashmiri) share is that they are typically analysed with a metrical unit at the right word edge.[21] Why would such languages be problematic for the Forward Learner? 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 analysable with a metrical unit at the right word edge. If this were the case, the problem is not with the generalisation 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.

6.8 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

[20] For example, *n*-gram-based learners cannot learn unbounded patterns, unless augmented with *a priori* projections (Hayes & Wilson 2008).

[21] According to Walker (2000), the Kashmiri data comes from Kenstowicz (1993), citing Bhatt (1989).

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 and 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 14 shows the suffix tree for the words with eight or fewer syllables which obey the stress pattern of Asmat.

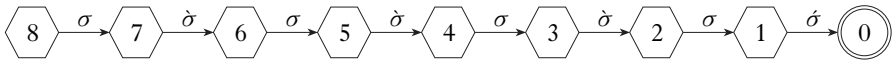


Figure 14
Suffix tree for Asmat words of eight syllables or fewer.

The structure of the suffix tree of this representation is very different from the structure in the prefix tree shown in Fig. 10. Yet both accept exactly the same (finite) set of words. Because they have different structures, the states in a suffix tree may have different neighbourhoods than the states in a prefix tree. Consequently, the generalisations acquired by merging states with the same neighbourhoods may be different. In the case of Asmat, the language obtained by merging same-neighbourhood states in the suffix tree is the same as the one returned by merging same-neighbourhood states in the prefix tree – it is the target language. But in the case of Lower Sorbian, the language obtained by merging same-neighbourhood states (in the suffix tree) is exactly the target language. In other words, this procedure avoids the overgeneralisation that the Forward Learner makes (again, see Heinz 2006 for details).

The learner which merges same-neighbourhood states in the suffix tree is called the Backward Neighbourhood Learner (BL). This learner identifies 96 of the 109 distinct patterns in the database, 82 of which the FL also identifies. As in the case of the FL, the error cases are ones of overgeneralisation. However, unlike the FL's errors, these occur predominantly in languages which are left-edge based.

Of course it is desirable to have a learner which does not know in advance whether to use a prefix or suffix tree. The FBL below is one such learner.

6.9 The Forward Backward Neighbourhood Learner

The FBL is very simple. Let M_{nd} be the function which maps an acceptor to the acceptor obtained by merging same-neighbourhood 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 M_{nd} to the prefix and suffix tree representations of the samples and intersects the results.

(14) *Algorithm 2: The Forward Backward Neighbourhood 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 $A = A_1 \times A_2$ and output acceptor A

The product ($\times$) of two acceptors A and B is also called the intersection of two machines, because this operation results in an acceptor which only accepts words accepted by both A and B (Hopcroft *et al*. 2001). In other words, $L(A \times B) = L(A) \cap L(B)$. This learner succeeds on 100 of the 109 patterns (413 of 422 languages), a considerable improvement over the Forward and Backward Learners. Appendix A provides these results, along with those of the BL and FL. The following discussion below addresses why the learner works, and compares the learnable and unlearnable patterns to the attested typology.

7 Discussion

7.1 Basic reasons why the FBL works

The reason the FBL succeeds in more cases than the FL and BL is simple: intersection preserves the robust generalisations. The robust generalisations are the ones made by merging states in *both* the prefix and suffix trees. Overgeneralisations that are made by the FL are not always made by merging same-neighbourhood states in the suffix tree. Consequently, such overgeneralisations do not survive the intersection process. Likewise, it is also true that overgeneralisations made by merging same-neighbourhood states in the suffix tree are not always made in the prefix tree.

For example, recall that the pattern obtained by the FL for Lower Sorbian is an overgeneralisation – words with four or more syllables are stressed in two ways: primary on the initial and secondary on the penult, or simply primary on the initial (e.g. both *σσὸσ* and *σσσσ* are generated). However, the pattern obtained by the BL is exactly the target pattern which requires secondary stress on the penult in words with four or more syllables. The intersection cuts out the overgeneralisation made by the FL.

However, it is the generalisation strategy itself – the merging of same-neighbourhood states – which is the real reason for the algorithm's success. Consider again the FL. By merging states with the same neighbourhood, the algorithm guarantees that its output is neighbourhood-distinct. Similarly, when the same-neighbourhood states are merged in the suffix tree, the resulting acceptor is neighbourhood-distinct. The learner – by merging same-neighbourhood states – generalises to neighbourhood-distinct patterns. Thus if people generalise similarly, it explains why nearly all stress patterns are neighbourhood-distinct.

There is one caveat, however. As explained in §4.4, the class of neighbourhood-distinct languages is not closed under intersection. Thus when the Forward Backward Neighbourhood Learner intersects the two acceptors obtained by merging same-neighbourhood states in the prefix and suffix trees, the resulting language is not guaranteed to be neighbourhood-distinct. Little is understood about what additional properties are necessary to ensure that neighbourhood-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 neighbourhood-distinct pattern for every pattern in the study (except Asheninca; Payne 1990).

7.2 Unlearnable unattested patterns

It is also interesting to note that most unattested patterns cannot be learned by the FBL. Intuitively, this follows from the fact that neither the FL nor BL can ever learn a non-neighbourhood-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 again the acceptor in Fig. 4, which generates the logically possible stress pattern which assigns stress to the initial syllable and then every fourth syllable. The reason that this pattern cannot be learned by the FBL is because states 2 and 3 have the same neighbourhood. Consequently, neither the FL nor the BL could ever arrive at this pattern by merging same-neighbourhood 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 overgeneralisation is made by both FL and BL learners, it survives the intersection process. Thus the result obtained by the FBL 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-neighbourhood states.

Albert Einstein is claimed to have said: 'Many of the things you can count, don't count. Many of the things you can't count, really count'. In the context of language learning, this quotation has new meaning, because 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.

It is not immediately obvious that the notion of locality ought to be sufficient for learning any stress patterns at all. This paper has shown that a particular formulation of an *a priori* notion of locality can make a significant contribution to language learning. Thus, at the very least, the FBL predicts that logically possible stress patterns like the quaternary one above should be 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 FBL fails to learn, which constitute a direct challenge to the viability of Hypothesis 2. Recall that the learner fails if the acceptor obtained does not generate *exactly* the same language as the target one. One question to keep in mind throughout this discussion is the extent of the difference between the obtained patterns and the target ones.

In every case the FBL fails because it overgeneralises. 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 ways. In other words, the learner allows a certain degree of optionality. This happens 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 (in Kelkar's analysis), are not canonically neighbourhood-distinct, and are discussed in §4.6.

Mingrelian (Klimov 2001) is a neighbourhood-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: 75–98), Negev Bedouin Arabic (Kenstowicz 1983) and Hindi (in the analysis of Fairbanks 1981) are not learnable by this learner, though they are neighbourhood-distinct. It is striking that these are precisely the patterns in the typology that have been analysed with extrametrical feet (Hayes 1995), suggesting that the FBL cannot identify in the limit from positive data the class of patterns describable with extrametrical feet.

Ashéninca (Payne 1990) and Pirahã (Everett 1988) are two other patterns which are neighbourhood-distinct but also beyond the reach of the FBL. These patterns are well-known prominence systems (Hayes 1995). However, I suspect the reason the FBL 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 FBL can learn only some patterns of this type (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.[22] 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. This ought to be investigated more closely in future research to obtain a better understanding.

Since nine of the stress patterns are not learned by the algorithm, we might 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 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 words of four syllables (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 quadrisyllabic 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.

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. In

[22] 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-neighbourhood-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 neighbourhoods. This is like the pattern in Yidiŋ, except that Yidiŋ is neighbourhood-distinct and FBL learnable, because of the distinction it maintains between heavy and light syllables.

Ashéninca (Payne 1990), the FBL predicts that words ending with a long vowel followed by three syllables with a high front vowel, like attested ['ma:kiriti] 'type of bee', could have two pronunciations: reported ['ma:kiriti], but also [ˌma:ki'riti]. According to Fairbanks, stress in Hindi falls on the initial syllable in disyllabic words. The only overgeneralisation 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 those described by previous researchers. 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 (i) in certain words there will be optionality and (ii) 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 have been overlooked (or in the case of secondary stress, difficult to detect) in earlier hypothesis formation.[23]

Not all the overgeneralisations made by the learner may be as plausible as the above discussion might suggest. For example, the only overgeneralisation made by the FBL when learning Kelkar's description of Hindi occurs in words of five syllables or longer. However, some of the optionally acceptable forms include no primary stress. Instead, secondary stress occurs where primary stress should fall (in addition to the positions where it is expected). The FBL acceptor obtained for Pirahã can place stress according to the Pirahã pattern or in some words optionally on the final syllable. 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). The overgeneralisations in Palestinian Arabic involve only words at least four syllables in length whose penult and antepenult are light, and in Cyrenaican Bedouin Arabic the overgeneralisation occurs in certain words of three syllables or more. In these cases, the degree of optionality appears greater than 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 neighbourhood-distinctness.

[23] We might also expect 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) and Seneca (Chafe 1977, Stowell 1979) then the FBL fails to learn these grammars. It fails because, in the actual grammars of Klamath and Seneca, the presence of secondary stress *distinguishes* the neighbourhoods of certain states of the prefix and/or suffix trees.

Finally, note that if the FBL is modified so that it only merges states with the same 2-2 neighbourhoods, then all overgeneralisation is eliminated – the FBL successfully learns every pattern in the typology.[24]

7.4 Learnable unattested patterns

The FBL can also learn many unattested patterns that are unnatural and not present in the typology. However, the learners developed here primarily examine the contribution that locality can make to learning. This contribution is significant – it is a sufficient property for learning to occur. However, in no way should we expect locality 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'. To my knowledge, no such stress pattern is attested. Whether or not humans can learn such a pattern is an open question, and, as far as I know, there is no experimental evidence bearing on it. However, even if it were shown that 'leftmost light, otherwise rightmost' is more difficult to learn than the more natural 'leftmost heavy, otherwise rightmost' pattern (providing evidence for our intuitions that 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 1990).

8 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 these iterative learners converge to the correct grammar.

I begin this discussion by pointing out one fact about the input samples described earlier. Making the samples consist of words from one to n syllables guarantees in prefix and suffix tree construction that many states fully realise their possible outgoing transitions. I will call a state in a prefix or suffix tree which has every possible incoming and outgoing path realised (as determined by the sampling language) SATURATED. As illustration, consider the prefix tree for the 'leftmost heavy, otherwise rightmost' words in Table I shown in Fig. 15, where shading indicates final states. Because of the kind of sample used to construct the trees, every non-terminal state in Fig. 15 is saturated.

When we consider iterative learners, it is not the case that all states up to a certain depth in the prefix and suffix trees will be saturated. For example, if the sample in Table I for the 'leftmost heavy, otherwise

[24] Though one undesirable consequence of 2-2 neighbourhood learning is that patterns describable with feet with four syllables are predicted to exist.

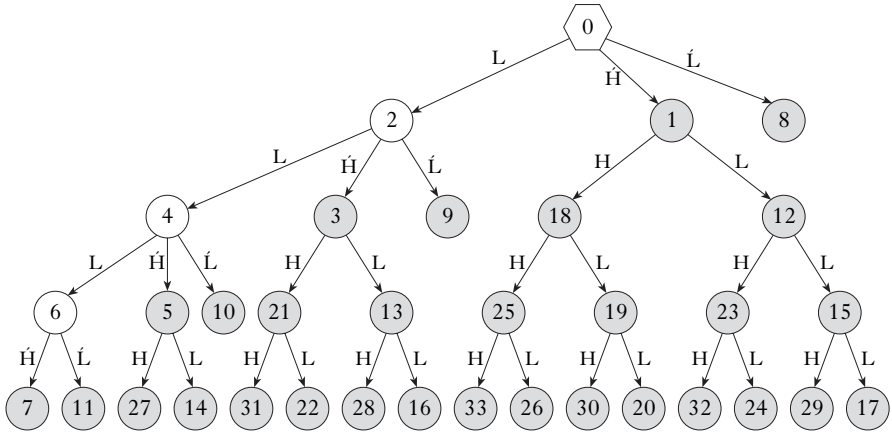


Figure 15

Prefix tree of the 'leftmost heavy, otherwise rightmost' words in Table I.

rightmost' pattern excluded word types ĤL, ĤLH, ĤLHH, ĤLHL, then the branch in the prefix tree from state 12 to 23 would be missing, and state 12 would be non-final. If LLĹ, LLĤ, LLĤL, LLĤH 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 and non-start, and each would have only one incoming transition, labelled L, and one outgoing transition, also labelled L. Such a merge would be an error, since the resulting pattern obtained would include forms like ĤLLĤ.

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. However, it can be shown that if the sample is sufficient (i.e. certain states in the trees are saturated), then same-neighbourhood 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 Neighbourhood Learner. Imagine that 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 (i) there is some time step t where the learner hypothesises a grammar generating the target pattern, (ii) for all future time steps, the learner's hypothesis stays the same and (iii) t can be obtained after some finite number of time steps.

We modify the FL 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-neighbourhood 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,[25] in which case the acceptor representing the smaller pattern is chosen.[26]

As an illustration, in the case above, suppose that the current hypothesis H the learner obtains is based on the sample described, and so incorrectly accepts words like ĹLLĹ. If the next word heard by the learner is ĹLH, then the updated prefix tree now distinguishes states 12 from 4, and they will not be merged. Since H does not accept ĹLH, the learner chooses the new hypothesis H' obtained by merging same-neighbourhood states in the updated prefix tree. Similarly, if the next word is ĹL, then state 12 is saturated, and the new hypothesis H'', which merges state 12 with 15 and 19 to make the right generalisation, will be selected.

It is easy to show that this iterative learner eventually converges to the target patterns the batch learner succeeds on, no matter the order of the presentation of the learner's input. This is because in these cases there is a finite sample such that the prefix tree contains states which are saturated up to a certain depth and that merging same-neighbourhood states in this prefix tree returns an acceptor equal to the target pattern.[27] Therefore, by the theorem in Appendix B, merging same-neighbourhood-states in the prefix tree built from any larger sample returns a language which is a superset of the target pattern. Thus, at this point, the iterative learner above rejects any new hypothesis and keeps the hypothesised grammar that generates the target language. Therefore we have exact identification in the limit.[28]

The incremental learner above requires maintaining the prefix tree in its memory; as such, it is not a memoryless online learner. However, the prefix tree is really a representation of the lexicon, which is needed independently. The only controversial aspect is the use of suffix trees, which

[25] One pattern is a subset of another provided the intersection of the complement of one acceptor and the other acceptor is empty.

[26] 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.

[27] 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).

[28] Admittedly, we are in the strange position of not knowing exactly what class of languages is being identified in the limit. See Heinz (2007, 2008) for early results towards this end.

provides an additional, unorthodox representation of the lexicon. A reviewer points out that right-to-left representations are undesirable from a processing point of view, which always proceeds temporally, i.e. from left to right. This raises a real question, which has not been investigated to my knowledge, which is how undeniably right-to-left patterns such as Asmat in (3b) are processed in time. What the FSA representation makes clear is that such patterns are more usefully dealt with in right-to-left terms (i.e. represented with reverse deterministic FSAs, obtained from merging states in suffix trees, etc.). This leaves open the question of how patterns like Asmat can be reconciled with left-to-right processing models.

9 Comparison to other learning models

Here I compare the FBL 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 FBL, each of these learning models was evaluated with respect to a typology of 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.

There are other learning models which tackle stress patterns, but because those models have different goals, I exclude them from comparison. For example, Pearl (2007) applies a stress learner designed to handle noise and lexical exceptions to English. While this work is interesting and insightful, it differs from the other learners above, which are evaluated according to some collected typology of languages, and which ignore noise and lexical exceptions in the input to their learners.

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, whether accuracy increases if longer words are admitted into the sample is unknown, but it is perfectly conceivable (and in my opinion quite likely).

Gupta & Touretzky (1994) present a perceptron with 19 stress patterns, of which it successfully learns 17. The training input consists of a sample of all words of seven syllables or fewer, and is presented to the perceptron at least 17 times. This is the smallest number that results in successful learning of any of the 19 patterns (e.g. the perceptron learned Latvian (Fennell & Gelsen 1980), 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 (Boas & Deloria 1941), 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 & Smolensky (2000) report twelve constraints, which yield 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. 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 ~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 the FBL, which has no *a priori* P&P parameters or OT constraints, is certainly 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 very much slower than the FBL, as it requires repeated presentations of words.[29]

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 neighbourhood-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.

10 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 is the set of 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, an examination of each stress pattern – represented by a finite-state acceptor – in two recent surveys (Bailey 1995,

[29] This should not necessarily be seen as a disadvantage. The perceptron does not have a memory to store words in the same way that these other learners do.

Gordon 2002), which yield a typological survey of 422 stress languages and 109 distinct stress patterns, revealed that 107 of them are tail- or head-canonically 1-1 neighbourhood-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 neighbourhood-distinct. Thus neighbourhood-distinctness approximates the attested typology in a non-trivial way. This leads to one hypothesis put forward in this paper, that all phonotactic patterns (of which stress patterns form a subclass) are canonically 1-1 neighbourhood-distinct.

Neighbourhood-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 Neighbourhood Learner, which merges same-state neighbourhoods 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-neighbourhood-distinct patterns, such as logically possible but unattested stress patterns that are describable with feet with four or more syllables. 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 learner's use of the reverse-deterministic suffix trees appears particularly suited for right-to-left based stress patterns. 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 overgeneralisations 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.

(i) In OT, what are the typological consequences of requiring constraints in Con to be neighbourhood-distinct?

(ii) Are there additional characterisations of neighbourhood-distinct patterns and languages learnable by the FBL, and what are they?

(iii) As the languages which contest Hypotheses 1 and 2 are more carefully investigated, do the challenges hold up?

(iv) As we enlarge the stress typology, do additional stress patterns of languages conform to Hypotheses 1 and 2, or not?

(v) Do adults or children learn neighbourhood-distinct patterns more easily than non-neighbourhood-distinct patterns?

In addition to these questions, I would like to mention two other avenues of research that appear fruitful. First, where do stress patterns fall in the Subregular Hierarchy? This is a hierarchy which categorises regular languages according to their inherent complexity, much like the Chomsky Hierarchy does for languages in general. 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 – coincide exactly to define this hierarchy. Some relevant recent results can be found in Edlefsen *et al.* (2008) and Graf (to appear).

More generally, Rogers & Pullum (2007) argue that the Subregular Hierarchy provides fertile ground for investigating 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 Rogers & Pullum. Artificial language learning experiments such as poverty of the stimulus experiments (Onishi *et al.* 2002, Chambers *et al.* 2003) appear capable of shedding light on this kind of question.

The second goal is to determine more precisely what kind of sample is needed to guarantee correct generalisation by the FBL. It is possible 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) provide an example of the kind of research that this last line of inquiry can lead to. Realising 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 underlying/surface pairs based on words found in the *Carnegie Mellon University English Pronouncing Dictionary*. 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 underlying/surface pairs). On the other hand, Gildea & Jurafsky show that the flapping rule is not inferred from the 50,000 underlying/surface pairs based on the dictionary.[30] 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. 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 will lead to new

[30] 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 & Jurafsky, which arguably offers the learner a richer linguistic environment than that which children are exposed to.

insights. In this case, it also remains an open question how these additional factors reduce the size of the necessary input sample for correct generalisation to occur.[31]

REFERENCES

Abrahamson, Arne (1968). Contrastive distribution of phoneme classes in Içuã Tupi. *Anthropological Linguistics* **10:6**. 11–21.

Albro, Daniel M. (2005). *Studies in computational Optimality Theory, with special reference to the phonological system 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**. 741–765.

Anthony, Martin & Norman Biggs (1992). *Computational learning theory*. Cambridge: Cambridge University Press.

Bailey, Todd M. (1995). *Nonmetrical constraints on stress*. PhD dissertation, University of Minnesota. (Stress System Database available (June 2009) at http://www.cf.ac.uk/psych/subsites/ssdb/.)

Barker, M. A. R. (1963). *Klamath dictionary*. Berkeley & Los Angeles: University of California Press.

Barker, M. A. R. (1964). *Klamath grammar*. Berkeley & Los Angeles: University of California Press.

Bhatt, Rakesh (1989). Syllable weight and metrical structure of Kashmiri. Ms, University of Illinois, Urbana.

Boas, Franz & Ella Deloria (1941). *Dakota grammar*. Washington: United States Government Printing Office.

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

Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **32**. 45–86.

Brame, Michael K. (1974). The cycle in phonology: stress in Palestinian, Maltese, and Spanish. *LI* **5**. 39–60.

Chafe, Wallace L. (1977). Accent and related phenomena in the Five Nations Iroquois languages. In Hyman (1977b). 169–181.

Chambers, Kyle E., Kristine H. Onishi & Cynthia Fisher (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition* **87**. B69–B77.

Chomsky, Noam (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam & Morris Halle (1965). Some controversial questions in phonological theory. *JL* **1**. 97–138.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Clark, Robin (1992). The selection of syntactic knowledge. *Language Acquisition* **2**. 83–149.

[31] Another line of inquiry raised by Gildea & 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?

Coleman, John (2005). *Introducing speech and language processing*. Cambridge: Cambridge University Press.
Crowhurst, Megan & Lev Michael (2005). Iterative footing and prominence-driven stress in Nanti (Kampa). *Lg* **81**. 47–95.
Denis, François, Aurélien Lemay & Alain Terlutte (2002). Some classes of regular languages identifiable in the limit from positive data. In Peter Adriaans, Henning Fernau & Menno van Zaanen (eds.) *Proceedings of the 6th International Colloquium on Grammatical Inference* (*ICGI*). Berlin: Springer. 63–76.
Donaldson, Bruce C. (1993). *A grammar of Afrikaans*. Berlin & New York: Mouton de Gruyter.
Dresher, B. Elan (1999). Charting the learning path: cues to parameter setting. *LI* **30**. 27–67.
Dresher, B. Elan & Jonathan D. Kaye (1990). A computational learning model for metrical phonology. *Cognition* **34**. 137–195.
Edlefsen, Matt, Dylan Leeman, Nathan Meyers, Nathaniel Smith, Molly Visscher & David Wellcome (2008). Deciding Strictly Local (SL) languages. In *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics* **6**. 6–75.
Eisner, Jason (1997a). Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual Meeting of the ACL and 8th EACL*. Madrid. 313–320.
Eisner, Jason (1997b). What constraints should OT allow? Handout from paper presented at the 71st Annual Meeting of the Linguistic Society of America, Chicago. Available as ROA-204 from the Rutgers Optimality Archive.
Eisner, Jason (1998). FootForm decomposed: using primitive constraints in OT. *MIT Working Papers in Linguistics* **31**. 115–143.
Ellison, T. Mark (1991). The iterative learning of phonological constraints. Ms, University of Western Australia.
Ellison, T. Mark (1992). *The machine learning of phonological structure*. PhD dissertation, University of Western Australia.
Ellison, T. Mark (1994). Phonological derivation in Optimality Theory. In *Proceedings of the 15th International Conference on Computational Linguistics* (*COLING*). Kyoto. 1007–1013.
Everett, Daniel (1988). On metrical constituent structure in Pirahã phonology. *NLLT* **6**. 207–246.
Fairbanks, Constance (1981). *The development of Hindi oral narrative meter*. PhD dissertation, University of Wisconsin, Madison.
Fennell, Trevor G. & Henry Gelsen (1980). *A grammar of modern Latvian*. 3 vols. The Hague: Mouton.
Fernau, Henning (2003). Identification of function distinguishable languages. *Theoretical Computer Science* **290**. 1679–1711.
Frank, Robert & Giorgio Satta (1998). Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics* **24**. 307–315.
Garcia, Pedro, Enrique Vidal & José Oncina (1990). Learning locally testable languages in the strict sense. In S. Arikawa, S. Goto, S. Ohsuga & T. Yokomori (eds.) *Algorithmic learning theory : 1st International Workshop*. 325–338.
Gerdemann, Dale & Gertjan van Noord (2000). Approximation and exactness in finite state optimality theory. In Jason Eisner, Lauri Karttunen & Alain Thériault (eds.) *Finite-state phonology : Proceedings of the 5th Workshop of the ACL Special Interest Group in Computational Phonology* (*SIGPHON*). Luxemburg. 34–45.
Gibson, Edward & Kenneth Wexler (1994). Triggers. *LI* **25**. 407–454.
Gildea, Daniel & Daniel Jurafsky (1996). Learning bias and phonological-rule induction. *Computational Linguistics* **22**. 497–530.

Gillis, Steven, Gert Durieux & Walter Daelemans (1995). A computational model of P&P: Dresher and Kaye (1990) revisited. In Maaike Verrips & Frank Wijnen (eds.) *Approaches to parameter setting*. Amsterdam: University of Amsterdam. 135–173.
Goedemans, Rob, Harry van der Hulst & Ellis Visch (eds.) (1996). *Stress patterns of the world*. Part 1: *Background*. The Hague: Holland Academic Graphics.
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: CSLI. 1–28.
Goldsmith, John & Jason Riggle (ms). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony.
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 Spenador, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.
Gordon, Matthew (2002). A factorial typology of quantity-insensitive stress. *NLLT* **20**. 491–552. Additional appendices available (June 2009) at http://www.linguistics.ucsb.edu/faculty/gordon/pubs.html.
Graf, Tomas (to appear). Comparing incomparable frameworks: a model theoretic approach to phonology. *Proceedings of the Penn Linguistics Colloquium*.
Greenberg, Joseph H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg. *Universals of language*. Cambridge, Mass.: MIT Press. 73–113.
Greenberg, Joseph H. (1978). Some generalisations concerning initial and final consonant clusters. In Joseph H. Greenberg (ed.) *Universals of human languages*. Vol. 2: *Phonology*. Stanford: Stanford University Press. 243–279.
Gupta, Prahlad & David Touretzky (1991). What a perceptron reveals about metrical phonology. In *Proceedings of the 13th 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**. 1–50.
Halle, Morris (1978). Knowledge unlearned and untaught: what speakers know about the sounds of their language. In Morris Halle, Joan Bresnan & George A. Miller (eds.) *Linguistic theory and psychological reality*. Cambridge, Mass.: MIT Press. 294–303.
Halle, Morris & Jean-Roger Vergnaud (1987). *An essay on stress*. Cambridge, Mass.: MIT Press.
Hammond, Michael (1986). The obligatory-branching parameter in metrical theory. *NLLT* **4**. 185–228.
Hammond, Michael (1987). Hungarian cola. *Phonology Yearbook* **4**. 267–269.
Harrison, Michael A. (1978). *Introduction to formal language theory*. Reading: Addison Wesley.
Hayes, Bruce (1995). *Metrical stress theory: principles and case studies*. Chicago: University of Chicago Press.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**. 379–440.
Heinz, Jeffrey (2006). Learning quantity insensitive stress systems via local inference. In Richard Wicentowski & Grzegorz Kondark (eds.) *Proceedings of the 8th Meeting of the ACL Special Interest Group in Computational Phonology*. New York City. 21–30.

Heinz, Jeffrey (2007). *The inductive learning of phonotactic patterns*. PhD dissertation, University of California, Los Angeles.
Heinz, Jeffrey (2008). Learning left-to-right and right-to-left iterative languages. In Alexander Clark, François Coste & Laurent Miclet (eds.) *Grammatical inference : algorithms and applications*. Berlin : Springer. 84–97.
Higuera, Colin de la (2005). A bibliographical study of grammatical inference. *Pattern Recognition* **38**. 1332–1348.
Higuera, Colin de la (in press). *Grammatical inference : learning automata and grammars*. Cambridge : Cambridge University Press.
Hopcroft, John E., Rajeev Motwani & Jeffrey D. Ullman (2001). *Introduction to automata theory, languages, and computation*. Boston : 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**. 313–359.
Hyman, Larry M. (1977a). On the nature of linguistic stress. In Hyman (1977b). 37–82.
Hyman, Larry M. (ed.) (1977b). *Studies in stress and accent*. Los Angeles : Department of Linguistics, University of Southern California.
Idsardi, William J. (1992). *The computation of prosody*. PhD dissertation, MIT.
Idsardi, William J. (2008). Calculating metrical structure. In Eric Raimy & Charles E. Cairns (eds.) *Contemporary views on architecture and representations in phonology*. Cambridge, Mass. : MIT Press. 191–212.
Jain, Sanjay, Daniel Osherson, James S. Royer & Arun Sharma (1999). *Systems that learn : an introduction to learning theory*. 2nd edn. Cambridge, Mass. : MIT Press.
Johnson, C. Douglas (1972). *Formal aspects of phonological description*. The Hague & Paris : Mouton.
Jones, W. E. (1971). Syllables and word-stress in Hindi. *Journal of the International Phonetic Association* **1**. 74–80.
Jurafsky, Daniel & James H. Martin (2000). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ : Prentice Hall.
Kager, René (1999). *Optimality Theory*. Cambridge : Cambridge University Press.
Kager, René (2007). Feet and metrical stress. In Paul de Lacy (ed.) *The Cambridge handbook of phonology*. Cambridge : Cambridge University Press. 195–227.
Kaplan, Ronald & Martin Kay (1994). Regular models of phonological rule systems. *Computational Linguistics* **20**. 331–378.
Karttunen, Lauri (1998). The proper treatment of optimality in computational phonology. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing*. Ankara : Bilkent University. 1–12.
Kearns, Michael & Umesh Vazirani (1994). *An introduction to computational learning theory*. Cambridge, Mass. : MIT Press.
Kelkar, Ashok (1968). *Studies in Hindi-Urdu*. Vol. 1 : *Introduction and word phonology*. Poona : Deccan College.
Kenstowicz, Michael (1983). Parametric variation and accent in the Arabic dialects. *CLS* **19**. 205–213.
Kenstowicz, Michael (1993). Peak prominence stress systems and Optimality Theory. In *Proceedings of the 1st International Conference of Linguistics and Chosun University*. Foreign Culture Research Institute, Chosun University, Kwangju, Korea. 7–22.
Kenstowicz, Michael (1994). *Phonology in generative grammar*. Cambridge, Mass. & Oxford : Blackwell.
Klimov, G. A. (2001). Megrelskii yazyk. In M. E. Alekseev (ed.) *Yazyki mira : Kavkazskie yazyki*. Moscow : Izdatelstvo Academia. 52–58.

Kobele, Gregory (2006). *Generating copies : an investigation into structural identity in language and grammar*. PhD dissertation, University of California, Los Angeles.
Koskenniemi, Kimmo (1983). *Two-level morphology : a general computational model for word-form recognition and production*. Helsinki: Department of General Linguistics, University of Helsinki.
Kracht, Marcus (2003). *The mathematics of language*. Berlin & New York: Mouton de Gruyter.
McCarthy, John J. (2003). OT constraints are categorical. *Phonology* **20**. 75–138.
McCarthy, John J. & Alan Prince (1986). *Prosodic morphology*. Ms, University of Massachusetts, Amherst & Brandeis University.
McNaughton, Robert & Seymour A. Papert (1971). Counter-free automata. Cambridge, Mass.: MIT Press.
Mairal, Ricardo & Juana Gil (eds.) (2006). *Linguistic universals*. Cambridge: Cambridge University Press.
Martin, Andrew (2007). *The evolving lexicon*. PhD dissertation, University of California, Los Angeles.
Michelson, Karin (1988). *A comparative study of Lake-Iroquoian accent*. Dordrecht: Kluwer.
Mitchell, T. F. (1975). *Principles of Firthian linguistics*. London: Longman.
Moreton, Elliott (2008). Analytic bias and phonological typology. *Phonology* **25**. 83–127.
Muggleton, Stephen (1990). *Inductive acquisition of expert knowledge*. Wokingham: Addison-Wesley.
Niyogi, Partha (2006). *The computational nature of language learning and evolution*. Cambridge, Mass.: MIT Press.
Nowak, Martin A., Natalia L. Komarova & Partha Niyogi (2002). Computational and evolutionary aspects of language. *Nature* **417**. 611–617.
Ohala, Manjari (1977). Stress in Hindi. In Larry Hyman (1977b). 327–338.
Oncina, José, Pedro García & Enrique Vidal (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**. 448–458.
Onishi, Kristine H., Kyle E. Chambers & Cynthia Fisher (2002). Learning phonotactic constraints from brief auditory experience. *Cognition* **83**. B13–B23.
Partee, Barbara H., Alice ter Meulen & Robert E. Wall (1990). *Mathematical methods in linguistics*. Dordrecht: Kluwer.
Payne, Judith (1990). Asheninca stress patterns. In Doris L. Payne (ed.) *Amazonian linguistics : studies in lowland South American languages*. Austin: University of Texas Press. 185–209.
Pearl, Lisa (2007). *Necessary bias in natural language learning*. PhD dissertation, University of Maryland, College Park.
Piattelli-Palmarini, Massimo (ed.) (1980). *Language and learning : the debate between Jean Piaget and Noam Chomsky*. Cambridge, Mass.: Harvard University Press.
Popper, Karl R. (1959). *The logic of scientific discovery*. New York: Basic Books.
Prince, Alan (1983). Relating to the grid. *LI* **14**. 19–100.
Prince, Alan (1990). Quantitative consequences of rhythmic organization. *CLS* **26:2**. 355–398.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory : constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.
Riggle, Jason (2004). *Generation, recognition, and learning in finite-state Optimality Theory*. PhD dissertation, University of California, Los Angeles.
Roark, Brian & Richard Sproat (2007). *Computational approaches to morphology and syntax*. Oxford: Oxford University Press.

Rogers, James & Geoffrey Pullum (2007). Aural pattern recognition experiments and the subregular hierarchy. In Marcus Kracht (ed.) *Proceedings of the 10th Mathematics of Language Conference*. University of California, Los Angeles. 1–7.
Sharma, Aryendra (1969). Hindi word-accent. *Indian Linguistics* **30**. 115–118.
Sharpe, Margaret C. (1972). *Alawa phonology and grammar*. Canberra: Australian Institute of Aboriginal Studies.
Shieber, Stuart M. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy* **8**. 333–343.
Sipser, Michael (1997). *Introduction to the theory of computation*. Boston: PWS Publishing.
Stabler, Edward P. (2009). Computational models of language universals: expressiveness, learnability and consequences. In Morten H. Christiansen, Chris Collins & Simon Edelman (eds.) *Language universals*. Oxford: Oxford University Press. 200–223.
Stowell, T. (1979). Stress systems of the world, unite! *MIT Working Papers in Linguistics* **1**. 51–76.
Tenenbaum, Josh (1999). *A Bayesian framework for concept learning*. PhD dissertation, MIT.
Tesar, Bruce (1998). An iterative strategy for language learning. *Lingua* **104**. 131–145.
Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.
Valiant, L. (1984). A theory of the learnable. *Communications of the ACM* **27**. 1134–1142.
Voorhoeve, Clemens L. (1965). *The Flamingo Bay dialect of the Asmat language*. The Hague: Nijhoff.
Walker, Rachel (2000). Mongolian stress, licensing and factorial typology. Ms, University of California, Santa Cruz. Available as ROA-172 from the Rutgers Optimality Archive.
Wilson, Colin (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* **30**. 945–982.
Yang, Charles (2000). *Knowledge and learning in natural language*. PhD dissertation, MIT.