Connectionist Models and Linguistic Theory: Investigations of Stress Systems in Language

Prahlad Gupta[*] and David S. Touretzky[†]

April 1993

CMU-CS-93-146

[*]Department of Psychology
[†]Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

To appear in *Cognitive Science*

Abstract

We question the widespread assumption that linguistic theory should guide the formulation of mechanistic accounts of human language processing. We develop a pseudo-linguistic theory for the domain of linguistic stress, based on observation of the learning behavior of a perceptron exposed to a variety of stress patterns. There are significant similarities between our analysis of perceptron stress learning and metrical phonology, the linguistic theory of human stress. Both approaches attempt to identify salient characteristics of the stress systems under examination without reference to the workings of the underlying processor. Our theory and computer simulations exhibit some strikingly suggestive correspondences with metrical theory. We show, however, that our high-level pseudo-linguistic account bears no causal relation to processing in the perceptron, and provides little insight into the nature of this processing. Because of the persuasive similarities between the nature of our theory and linguistic theorizing, we suggest that linguistic theory may be in much the same position. Contrary to the usual assumption, it may not provide useful guidance in attempts to identify processing mechanisms underlying human language.

The second author was supported in part by grants from Hughes Aircraft Corporation and by the Office of Naval Research under contract number N00014-86-K-0678.

Keywords: cognitive science, language, connectionist models, linguistic theory, metrical phonology, stress systems, perceptron learning.

Contents

To appear in *Cognitive Science*

Connectionist Models and Linguistic Theory: Investigations of Stress Systems in Language

Prahlad Gupta and David S. Touretzky

The fundamental questions to be answered about language are, as Noam Chomsky has pointed out: (1) What is the system of knowledge in the mind/brain of the speaker of English or Spanish or Japanese? (2) how does this knowledge come into being? (3) how is this knowledge used? and (4) what are the physical mechanisms serving as the basis for this system of knowledge, and for its use? (Chomsky, 1988, p.3)

Linguistic theory is viewed, in the Chomskyan tradition, as illuminating the nature of mind, and as guiding the search for lower level processing. Thus, according to Chomsky, to the extent that the linguist can provide answers to the first three questions above, the brain scientist can begin to explore answers to the fourth question, i.e., can begin to explore the physical mechanisms that exhibit the properties revealed in the linguist's abstract theory. In the absence of answers to the first three questions, brain scientists do not know what they are searching for (Chomsky, 1988, p.6).

Chomsky distinguishes three levels of analysis in linguistic inquiry: observation, description, and explanation; these are the tasks for linguistics (Chomsky, 1988, p.60-62). Explanation is via the formulation of Universal Grammar, which is viewed as providing a genuine explanation of the phenomena (Chomsky, 1988, p.62). Universal Grammar reveals the principles of the mind (Chomsky, 1988, p.91), and thus constitutes the abstract theory that is needed to guide the search for actual brain mechanisms.

This view can be stated more generally as the claim that the abstract formulations of linguistic theory are the appropriate framework within which to organize the search for the lower-level processing mechanisms of language. Our aim in this paper is to draw attention to the fact that this is an *assumption*, and one with important consequences. To the extent that this claim is true, it suggests that formulation of accounts of language processing, as well as the search for the neural substrates of language, should be guided by the constructs of linguistic theory. To the extent that this claim is false, the use of linguistic constructs as the basis for processing accounts and neural investigation may be misguided, and misleading.

Our goal, then, is to examine whether it is necessarily true that the constructs of linguistic theory play a valuable role in formulation of accounts of lower-level processing. In this paper, we will develop a *pseudo-linguistic theory*[1] for the domain of linguistic stress, based on observation of the learning behavior of a perceptron exposed to a variety of stress patterns. We will then compare this theory with *metrical phonology*, the actual linguistic theory of stress, showing how our theory exhibits many of the kinds of regularities and predictions of metrical theory. Thus, our theory will be analogous, in important ways, to metrical theory. It predicts, for example, that certain stress systems will be unlearnable. We will then analyze the behavior of the underlying perceptron model in terms of its actual processing. This will show that our theoretical predictions of non-learnability bear no relation to the low-level, causal explanation we develop as part of the processing account.

In short, we will demonstrate that a lower-level account may not be isomorphic (or even similar) to a higher-level account, and that under these circumstances, the higher-level account may provide little insight into the underlying phenomena. Furthermore, since the nature and the mode of development of our higher-level account are closely analogous to the nature and development of bodies of linguistic theory, this constitutes a strong argument against assuming that the high-level accounts of linguistic theory will necessarily illuminate lower-level processing mechanisms, and against assuming that linguistic theory is the appropriate conceptual framework within which to organize the search for these mechanisms. Such arguments have of course been made before (Pylyshyn, 1973; Rumelhart and McClelland, 1987; Smolensky, 1988). However, the present work substantiates the argument with a concrete demonstration of how an analytical framework that bears a close resemblance to linguistic theory can fail to have an analogue in the underlying processing mechanisms.

The organization of the paper is as follows. The first section provides some background information about the

[1] We call our theory *pseudo*-linguistic only because it is based on data generated by the perceptron rather than humans.

domain of linguistic stress, while the second section describes our perceptron simulations of stress learning. The third section develops our pseudo-linguistic theory, and discusses correspondences between aspects of the perceptron simulations, our theory, and metrical theory. In the fourth section, we show that our theory in reality offers only an abstract account that is unrelated to low-level processing characteristics. Having shown that a high-level account need not illuminate underlying processing, we turn in the fifth section to a consideration of some other reasons why the analyses of metrical theory might not be appropriate primitives for a processing model. In view of all this, we conclude that the Chomskyan assumption that linguistic theory should guide investigation of processing mechanisms may be unwarranted.

1. BACKGROUND: STRESS SYSTEMS IN LANGUAGE

1.1. Motivation for Choice of Domain

In order to examine a subsystem of linguistic theory in terms of its implications for lower-level processing, a prerequisite is, of course, that the theory be clearly specified. According to Dresher & Kaye (1990), stress systems are an attractive domain for investigation because: (a) the linguistic theory is well-developed, so that compared with syntax, there is a relatively complete description of the observed phenomena, and (b) stress systems can be studied relatively independently of other aspects of language (Dresher and Kaye, 1990, p. 138). We therefore chose to construct our pseudo-linguistic theory with respect to linguistic stress, since it is regarded as a well-defined domain for which theoretical analyses provide good coverage.

1.2. Evolution of the Linguistic Theory

The analysis of stress has evolved through a number of phases: (I) Linear analyses presented stress as a phonemic feature of individual vowels, with different levels of stress representing different levels of absolute prominence. This approach was taken in Trager & Smith (1951), and culminated in Chomsky & Halle's seminal *The Sound Pattern of English* (1968). (II) Metrical theory, as developed by Liberman & Prince (Liberman, 1975; Liberman and Prince, 1977), introduced both a non-linear analysis of stress patterns (in terms of *metrical trees*), and the treatment of stress as a *relative* property rather than an *absolute* one; however, the *stress* feature was retained in the analysis. (III) In subsequent developments (Prince, 1976; Selkirk, 1980), reliance on this feature was eliminated by incorporation of the idea that subtrees of metrical trees had an independent status (*metrical feet*), so that stress assignment rules could make reference to them. (IV) The positing of internal structure for syllables (Vergnaud and Halle, 1978; McCarthy, 1979a; McCarthy, 1979b) provided a means of distinguishing light and heavy syllables, a distinction to which stress patterns are widely sensitive, but which had been problematic under previous analyses. (V) An analysis of metrical tree geometries (Hayes, 1980) provided an account of many aspects of stress systems in terms of a small number of *parameters*.

Through the development of metrical theory, there has been debate over whether the auto-segmental representations for stress are *metrical trees* only (Hayes, 1980), *metrical grids* only (Prince, 1983; Selkirk, 1984), or some combination of the two (Liberman, 1975; Liberman and Prince, 1977; Hayes, 1984a; Hayes, 1984b; Halle and Vergnaud, 1987a; Halle and Vergnaud, 1987b).

1.3. Syllable Structure

A syllable is analyzed as being comprised of an *onset*, which contains the material before the vowel, and a *rime*. The rime is comprised of a *nucleus*, which contains the vocalic material, and a *coda*, which contains any remaining (non-vocalic) material. For further discussion, see Kaye (1989, pp. 54-58).

A syllable may be *open* (it ends in a vowel); or *closed* (it ends in a consonant). In terms of syllable structure, an open syllable has a *non-branching* rime (the rime has a nucleus, but not a coda), and a closed syllable has a *branching* rime (the rime has both a nucleus and a coda).

In many languages, stress tends to be placed on certain *kinds* of syllables rather than on others; the former are termed *heavy* syllables, and the latter *light* syllables. What counts as a heavy or a light syllable may differ across languages in which such a distinction is present, but, most commonly, a heavy syllable is one that can be characterized as having a branching rime, and a light syllable can be characterized as having a non-branching rime (Goldsmith, 1990, p. 113). Languages that involve such a distinction (between heavy and light syllables, i.e., between the *weights* of syllables) are termed *quantity-sensitive*, and languages that do not, *quantity-insensitive*. Note that in quantity-insensitive languages syllables can occur both with and without branching rimes, but the distinction between these kinds of syllables has no relevance for the placement of stress.

1.4. Metrical Phonology

There seems to be theoretical agreement that stress patterns are sensitive to information about syllable structure, and in particular, to the structure of the syllable *rime*, and not the syllable *onset*. We follow this assumption[2]. Thus rime structure is taken to be the basic level at which accounts of stress systems are formulated. (For an overview of metrical theory, see Goldsmith (1990, chapter 4), Kaye (1989, p. 139-145), van der Hulst & Smith (1982), or Dresher & Kaye (1990, p. 1-8)). Stress patterns are controlled by metrical structures built on top of rime structures. The version of metrical structure adopted here is *metrical feet*. We assume the *parameters* formulated by Dresher & Kaye (1990, p. 142):

(P1) The word-tree is strong on the [Left/Right]
(P2) Feet are [Binary/Unbounded]
(P3) Feet are built from the [Left/Right]
(P4) Feet are strong on the [Left/Right]
(P5) Feet are Quantity-Sensitive (QS) [Yes/No]
(P6) Feet are QS to the [Rime/Nucleus]
(P7) A strong branch of a foot must itself branch [No/Yes]
(P8) There is an extrametrical syllable [Yes/No]
(P9) It is extrametrical on the [Left/Right]
(P10) A weak foot is defooted in clash [No/Yes]
(P11) Feet are non-iterative [No/Yes]

As an example of the application of these parameters, consider the stress pattern of Maranungku, in which primary stress falls on the first syllable of the word and secondary stress on alternate succeeding syllables. Figure 1 shows an abstract representation of a six-syllable word, with each syllable represented as σ. The assignment of stress is characterized as follows. Binary, quantity-insensitive, left-dominant feet are constructed iteratively from the left edge of the word. Each foot has a "strong" and a "weak" branch (labeled "S" and "W," respectively, in the figure). The strong, or dominant branch assigns stress to the syllable it dominates. Since the feet are left-dominant, odd-numbered syllables are assigned stress. Over the roots of these *metrical feet*, a left-dominant *word-tree* is constructed, which assigns stress to the structure dominated by its leftmost branch. The third and fifth syllables are each dominated by the dominant branch of one metrical structure (a foot), while the first syllable is dominated by the dominant branches of two structures (a foot, and the word-tree). Even-numbered syllables are dominated only by non-dominant branches of feet. The result is that even-numbered syllables receive no stress; the third and fifth syllables receive one degree of stress (secondary stress); and the first syllable receives two degrees of stress (primary stress.) The parameter settings characterizing Maranungku are: [P1 Left], [P2 Binary], [P3 Left], [P4 Left], [P5 No], [P7 No], [P8 No], [P10 No], [P11 No]. Parameters P6 and P9 do not apply because of the settings of parameters P5 and P8, respectively.

1.5. Principles and Parameters

Metrical theory illustrates the *principles and parameters* approach to language, one of whose central hypotheses is that language learning proceeds through the discovery of appropriate parameter settings. Every possible human language can be characterized in terms of parametter settings; once these settings are determined, the nature of structure-sensitive operations and the structures on which they operate is known, so that the details of language processing are

---

[2]See, for example, Dresher & Kaye (1990, p. 141) or Goldsmith (1990, p. 170). Note, however, that some researchers have presented evidence that onsets may in fact be relevant to the placement of stress (Davis, 1988; Everett and Everett, 1984).

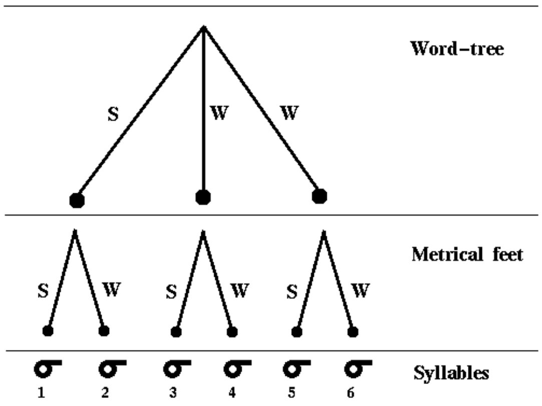

Figure 1: Metrical structures for a six-syllable word in Maranungku.

automatically determined (at an abstract level). Subsequently, the assignment of stress in the actual *production* or *processing* of language is assumed to involve neural processes that correspond quite directly with the abstract process of application of these parameter settings as guidelines to the construction and manipulation of metrical feet.

1.6. Previous Computational Models of Stress Learning

Computational models of stress systems in language have been developed by Dresher & Kaye (1990) and by Nyberg (1990, 1992). The focus of these models is on the learning of the *parameters* specified by metrical theory; they therefore take as a starting point the constructs of that theory, and incorporate its assumptions. What they add to the linguistic theory is what Dresher & Kaye term a *learning theory*: a specification of how data the language learner encounters in its environment are to be used to set parameters. The following features characterize these models: (1) they assume the existence of processes explicitly corresponding to the linguistic notion of parameter setting; (2) they propose a *learning* theory as an account of that parameter-setting process; (3) they assume that the process of *production* (i.e., of producing appropriate stress contours for input words after learning has occurred) involves explicit representational structures and structure-sensitive operations directly corresponding to metrical-theoretic trees and operations on those trees; (4) they assume no necessary relationship between the processing mechanisms involved in *learning* vs. those involved in *production*. That is, learning is accomplished by supplying values for the parameters defined by metrical theory; these values then form a knowledge base for stress assignment, whose processing involves, for example, the construction of binary trees from right to left – an operation having no necessary correspondence with those by which the parameter values were acquired.

The above models exemplify the assumption we are interested in examining in the present work: that abstract linguistic formulations are the appropriate starting point for processing models.

Processing in the perceptron will differ from the Dresher & Kaye and Nyberg models with regard to the above noted characteristics: (1) there is no explicit incorporation of parameters in the perceptron model; (2) the *learning theory* employed consists of one of the general learning algorithms common in connectionist modeling, and is not an account of parameter-setting; (3) the process of production does not involve explicitly structured representations in the classical sense; (4) the processing mechanisms and structures involved in production are essentially the same as those involved in learning. Thus, to the extent that we are successful in constructing an abstract, pseudo-linguistic theory that predicts behavior of the perceptron model, it will not be merely by virtue of having "built in" the constructs of metrical theory to begin with.

2. OVERVIEW OF MODEL AND SIMULATIONS

2.1. Nineteen Stress Systems

Nine quantity-insensitive (QI) languages and ten quantity-sensitive (QS) languages were examined in our experiments. The data, summarized in Table 1, were taken primarily from Hayes (1980). Note that the QI stress patterns of Latvian & French, Maranungku & Weri, Lakota & Polish, and Paiute & Warao are mirror images of each other. The QS stress patterns of Malayalam & Yapese, Ossetic & Rotuman, Eastern Permyak Komi & Eastern Cheremis, and Khalka Mongolian & Aguacatec Mayan are also mirror images[3].

2.2. Structure of the Model

In separate experiments, we taught a perceptron to produce the stress pattern of each of the nineteen languages. The domain was limited to single words, as in the previous learning models of metrical phonology developed by Dresher & Kaye and Nyberg. Again as in the other models, the effects of morpho-syntactic information such as lexical category were ignored, and the simplifying assumption was made that the only relevant information about syllables was their weight.

Two input representations were used. In the *syllabic* representation, used for QI patterns only, a syllable was represented as a [1, 1] vector, and [0, 0] represented no syllable. In the *weight-string* representation, which was necessary for QS languages, the input patterns used were [1, 0] for a light syllable, [0, 1] for a heavy syllable, and [0, 0] for no syllable. For stress systems with up to two levels of stress, the output targets used in training were 1.0 for primary stress, 0.5 for secondary stress, and 0 for no stress. For stress systems with three levels of stress, the output targets were 1.0 for primary stress, 0.6 for secondary, 0.35 for tertiary, and and 0 for no stress. The input data set for all stress systems consisted of all word-forms of up to seven syllables. With the syllabic input representation there are seven of these, and with the weight-string input representation, there are 254 distinct patterns.[4] The perceptron's input array was a buffer of 13 syllables; each word was processed one syllable at a time by sliding it through the buffer (see Figure 2). The desired output at each step was the stress level of the middle syllable of the buffer. Connection weights were adjusted at each step using the back-propagation learning algorithm[5] (Rumelhart et al., 1986). One *epoch* consisted of one presentation of the entire training set. The network was trained for as many epochs as necessary to ensure that the stress value produced by the perceptron was within 0.1 of the target value, for each syllable of the word, for all words in the training set. A *learning rate* of 0.05 and *momentum* of 0.90 were used in all simulations. Initial weights were uniformly distributed random values in the range ± 0.5. Each simulation was run at least three times, and the learning times averaged.

2.3. Training the Perceptron

To enable precise description, let the input buffer be regarded as a 2×13 array as shown in Figure 2. Let the 7th, i.e. center, column be numbered 0, let the columns to the left of the center be numbered negatively -1 through -6 going outwards from the center, and the columns to the right of the center be numbered positively +1 through +6, going outwards from the center.

---

[3] Descriptions somewhat more complex than ours have been reported for Polish (Halle and Vergnaud, 1987a, pp. 57-58) and Malayalam (Hayes, 1980, p. 66, 109). However, this does not detract from our discussion in any way, as stress systems corresponding to our simplifications are reported to exist: Swahili (Halle and Clements, 1983, p.17) and Gurkhali (Hayes, 1980, p.66), corresponding to Polish and Malayalam, respectively.

[4] In practice, we used weight string training sets in which there were an equal number of input patterns of each length. Thus, there was one instance of each of the 128 (= 2^7) seven-syllable patterns, and 64 instances of each of the two monosyllabic patterns. This length balancing was necessary for certain languages for successful training. We do not know how well it agrees with the actual frequency of forms in these languages, though as a general rule high-frequency forms are probably shorter.

[5] Note that although the architecture of the model is two-layered with a single output unit, as in a simple perceptron, we used the back-propagation algorithm (BP) rather than the Widrow-Hoff algorithm (WH); (Widrow and Hoff, 1960). BP adds to WH the scaling of the error for each output unit by a function (the derivative) of the activation of that output unit, and thus performs a more sensitive and example-tuned weight adjustment than WH. Note that BP and WH algorithms for two layers are guaranteed by the perceptron convergence procedure to be equivalent in terms of learning capabilities, *for binary-valued outputs*. However, outputs in the present simulations are not always binary; they sometimes take on intermediate values. As a result, the different computation of the error term in BP turned out to provide better learning than WH.

Table 1: Stress patterns: Description and example stress assignment.

REF	LANGUAGE	DESCRIPTION OF STRESS PATTERN	EXAMPLES
Quantity-Insensitive Languages:			
L1	Latvian	Fixed word-initial stress.	$S^1S^0S^0S^0S^0S^0S^0$
L2	French	Fixed word-final stress.	$S^0S^0S^0S^0S^0S^0S^1$
L3	Maranungku	Primary stress on first syllable, secondary stress on alternate succeeding syllables.	$S^1S^0S^2S^0S^2S^0S^2$
L4	Weri	Primary stress on last syllable, secondary stress on alternate preceding syllables.	$S^2S^0S^2S^0S^2S^0S^1$
L5	Garawa	Primary stress on first syllable, secondary stress on penultimate syllable, tertiary stress on alternate syllables preceding the penult, no stress on second syllable.	$S^1S^0S^0S^3S^0S^2S^0$
L6	Lakota	Primary stress on second syllable.	$S^0S^1S^0S^0S^0S^0S^0$
L7	Polish	Primary stress on penultimate syllable.	$S^0S^0S^0S^0S^0S^1S^0$
L8	Paiute	Primary stress on second syllable, secondary stress on alternate succeeding syllables.	$S^0S^1S^0S^2S^0S^2S^0$
L9	Warao	Primary stress on penultimate syllable, secondary stress on alternate preceding syllables.	$S^0S^2S^0S^2S^0S^1S^0$
Quantity-Sensitive Languages:			
L10	Koya	Primary stress on first syllable, secondary stress on heavy syllables. (Heavy = closed syllable or syllable with long vowel.)	$L^1L^0L^0H^2L^0L^0L^0$ $L^1L^0L^0L^0L^0L^0L^0$
L11	Eskimo	(Primary) stress on final and heavy syllables. (Heavy = closed syllable.)	$L^0L^0L^0H^1L^0L^0L^1$ $L^0L^0L^0L^0L^0L^0L^1$
L12	Malayalam	Primary stress on first syllable except when first syllable light and second syllable heavy. (Heavy = long vowel.)	$L^1L^0H^0L^0L^0L^0$ $L^0H^1L^0H^0L^0L^0L^0$
L13	Yapese	Primary stress on last syllable except when last is light and penultimate heavy. (Heavy = long vowel.)	$L^0L^0L^0H^0L^0L^0L^1$ $L^0H^0L^0H^0L^0H^1L^0$
L14	Ossetic	Primary stress on first syllable if heavy, else on second syllable. (Heavy = long vowel.)	$H^1L^0L^0H^0L^0L^0L^0$ $L^0L^1L^0L^0L^0L^0L^0$
L15	Rotuman	Primary stress on last syllable if heavy, else on penultimate syllable. (Heavy = long vowel.)	$L^0L^0L^0H^0L^0L^0H^1$ $L^0L^0L^0L^0L^0L^1L^0$
L16	Komi	Primary stress on first heavy syllable, or on last syllable if none heavy. (Heavy = long vowel.)	$L^0L^0H^1L^0L^0H^0L^0$ $L^0L^0L^0L^0L^0L^0L^1$
L17	Cheremis	Primary stress on last heavy syllable, or on first syllable if none heavy. (Heavy = long vowel.)	$L^0L^0H^0L^0L^0H^1L^0$ $L^1L^0L^0L^0L^0L^0L^0$
L18	Mongolian	Primary stress on first heavy syllable, or on first syllable if none heavy. (Heavy = long vowel.)	$L^0L^0H^1L^0L^0H^0L^0$ $L^1L^0L^0L^0L^0L^0L^0$
L19	Mayan	Primary stress on last heavy syllable, or on last syllable if none heavy. (Heavy = long vowel.)	$L^0L^0H^0L^0L^0H^1L^0$ $L^0L^0L^0L^0L^0L^0L^1$

Note. Examples are of stress assignment in seven-syllable words. Primary stress is denoted by the superscript 1 (e.g., S^1), secondary stress by the superscript 2, tertiary stress by the superscript 3, and no stress by the superscript 0. S indicates an arbitrary syllable, and is used for the QI stress patterns. For QS stress patterns, H and L are used to denote Heavy and Light syllables, respectively.

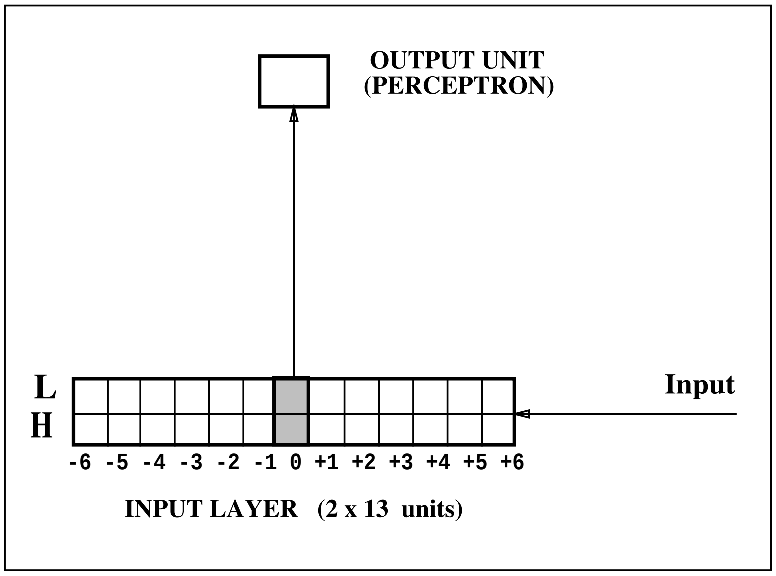

Figure 2: Perceptron model used in simulations.

As an example of processing, suppose that the next word in the training set for some quantity-insensitive language is the four-syllable pattern *[S S S S]*, and the associated stress contour (i.e., target), is *[0 0 0 1]*, indicating that the final syllable receives stress, and all other syllables are unstressed[6]. The first syllable enters the rightmost element of the input buffer (element +6). This is the first "time step" of processing. At the next time step, the first syllable (in the rightmost input buffer position, i.e., element +6) is shifted left to buffer element +5, and the second syllable enters the input buffer in element +6.

After two further time steps, elements +3, +4, +5, and +6 of the input buffer contain the four syllables of the current word. At the next two time steps, the leftward flow of syllables through the input buffer continues, until at time step 6, the word's four syllables are in elements +1 through +4. At time step 7, the first syllable moves into element 0 of the input buffer, and the four syllables of the word occupy elements 0 through +3 of the buffer. At this time step, training of the network occurs for the first time: the perceptron is trained to associate the pattern in its input buffer (the four syllables of the word, in buffer elements 0 through +3) with the target stress level for the syllable currently in buffer element 0. At the next time step, the syllables of the word are shifted left again, and occupy buffer elements -1 through +2. The network is trained to associate this pattern with the target stress level for the second syllable, which is now in element 0. Similarly, at the next two time steps, the perceptron is trained to produce the stress levels appropriate for the third and fourth syllables of the word, as they come to occupy element 0 of the input buffer.

At this point, the buffer is flushed (all elements are set to zero), and one "trial" is over. The next word in the training set can now enter, beginning the next trial.

Thus the processing of one word, syllable by syllable, constitutes one *trial*. One pass through all the words in the training set constitutes an *epoch*.

It should be noted that sequential processing of syllables is not a necessary part of this model. Exactly the same effect can be obtained by a parallel scheme of thirteen perceptron-like units whose weight vectors are tied together by *weight-sharing* (Hertz et al., 1991, p. 140). All thirteen units would learn in unison, and words could be processed in one parallel step.

[6]To simplify discussion here, each syllable is represented as an "S" token, rather than as an "H" or "L". For quantity-insensitive systems, this information suffices to determine placement of stress.

Table 2: Preliminary analysis of learning times for QI stress systems.

IPS	SCA	Alt	MPS	MSL	QI LANGUAGES	REF	EPOCHS (syllabic)
0	0	0	0	0	Latvian	L1	17
					French	L2	16
0	0	1	0	1	Maranungku	L3	37
					Weri	L4	34
0	1	1	0	1	Garawa	L5	165
1	0	0	0	0	Lakota	L6	255
					Polish	L7	254
1	0	1	0	1	Paiute	L8	**
					Warao	L9	**

Note. These learning times were obtained using the *syllabic* input representation. IPS=Inconsistent Primary Stress; SCA=Stress Clash Avoidance; Alt=Alternation; MPS=Multiple Primary Stresses; MSL=Multiple Stress Levels. Symbols L1-L9 refer to Table 1.

3. PERCEPTRON STRESS LEARNING: A PSEUDO-LINGUISTIC ACCOUNT

In this section, we analyze the characteristics of the various stress patterns to which the perceptron is exposed, and relate them to the perceptron's learning behavior, with a view to developing a *pseudo-linguistic theory*. In the same way that metrical phonology attempts to provide an account of linguistic stress in humans, our pseudo-linguistic theory will attempt to provide an account of the learning of stress in a perceptron.

Below, we first present an analysis of factors affecting learnability for the QI stress systems. We develop an analytic scheme that enables us both to characterize the patterns and to predict the ease or difficulty of their learning. We then include QS systems and show that one analytical framework can take into account both QI and QS systems. This overall framework constitutes our pseudo-linguistic theory.

3.1. Learnability of QI Systems

We begin by noting that learning times[7] differ considerably for {Latvian, French}, {Maranungku, Weri}, {Lakota, Polish} and Garawa, as shown in the last column of Table 2. Moreover, Paiute and Warao were unlearnable with this model[8].

Examination of the inherent features of these stress patterns suggests various factors as being relevant to learning:

Alternation of stresses (as opposed to a single stress) is suggested by the difference between learning times for {Latvian, French} and {Maranungku, Weri}, which also suggest that the *number of stress levels* may be relevant.

Recall from Table 1 that, in Garawa, primary stress is placed on the first syllable, secondary stress on the penultimate syllable, and tertiary stress on alternate syllables preceding the penultimate, but that no stress appears on the second syllable. The primary, secondary and tertiary stress patterns potentially lead to stress appearing on both the first and the second syllables; however, this is avoided (stress is never placed on the second syllable). This exemplifies the tendency in human languages to avoid the appearance of stress on adjacent syllables. The greater learning time for Garawa suggests that such *stress clash avoidance* is computationally expensive.

---

[7] These are learning times with the *syllabic* input representation.

[8] A case could perhaps be made that monosyllables are special (Elan Dresher, personal communication), so that it might be plausible to assume that there are no stressed monosyllables in the data. We investigated this possibility in two ways. First, if monosyllables are treated as receiving no stress, then the stress patterns of Paiute and Warao become learnable within the current architecture: it takes 45 epochs for the network to learn the stress pattern of Paiute, and 36 epochs for Warao. Second, if monosyllables are removed from the data set, the learning times are 44 epochs for Paiute, and 46 epochs for Warao.

Note also that both stress patterns *were* learnable, even with the stressed monosyllables, using either a two-layer architecture with two output units (targets were [0, 0] for no stress, [1, 0] for secondary stress, and [1, 1] for primary stress), or a three-layer architecture. We present the finding of non-learnability within the present model in order to maintain a consistency of analysis in one model, and since the two-layer model with a single output unit facilitates analysis of connection weights and their contribution to processing.

Table 3: Descriptions of hypothetical QI stress patterns.

REF	LANGUAGE	DESCRIPTION OF STRESS PATTERN
h1	Latvian2stress	Main stress on first syllable, secondary on second
h2	Latvian3stress	Main stress on first, secondary on second, tertiary on third syllable
h3	French2stress	Main stress on final, secondary on antepenult
h4	French3stress	Main stress on final, secondary on penult, tertiary on antepenult
h5	Latvian2edge	Main stress on first and last syllables
h6	Latvian2edge2stress	Main stress on first and last, secondary on antepenult
h7	Maranungku3stress	Main stress on first, secondary on penult, alternate preceding tertiary and secondary stresses
h8	Weri3stress	Main stress on last, secondary on antepenult, alternate preceding tertiary and secondary stress
h9	Latvian2edge2stress-alt	Main stress on first, secondary on penult and alternate preceding syllables
h10	Garawa-SC	Main stress on first, secondary on penult, tertiary on alternate preceding syllables
h11	Garawa2stress-SC	Main stress on first, secondary on penult and alternate preceding syllables
h12	Maranungku1stress	Main stress on first and alternate succeeding syllables
h13	Weri1stress	Main stress on last and alternate preceding syllables
h14	Latvian2edge-alt	Main stress on first and last and alternate preceding syllables
h15	Garawa1stress-SC	Main stress on first, and penult and alternate preceding syllables
h16	Latvian2edge2stress-1alt	Main stress on first, and antepenult and alternate preceding syllables, secondary on final
h17	Garawa-non-alt	Main stress on first, secondary on penult, tertiary on ante-antepenult, no stress on second
h18	Latvian3stress2edge-SCA	Main stress on first, secondary on last, tertiary on antepenult, no stress on second
h19	Latvian2edge-SCA	Main stress on first and last but no stress on second
h20	Latvian2edge2stress-SCA	Main stress on first and last, secondary on antepenult, no stress on second
h21	Garawa2stress	Main stress on first, secondary on penult and alternate preceding syllables, no stress on second
h22	Latvian2edge2stress-alt-SCA	Main stress on first, secondary on last and alternate preceding syllables, no stress on second
h23	Garawa1stress	Main stress on first, and penult and alternate preceding syllables, but no stress on second
h24	Latvian2edge-alt-SCA	Main stress on first, and last and alternate preceding syllables, no stress on second
h25	Latvian2edge2stress-1alt-SCA	Main stress on first, and antepenult and alt preceding syllables, secondary on last, but no stress on second
h26	Lakota2stress	Main stress on second, secondary on penult
h27	Lakota2edge	Main stress on second and penult syllables
h28	Lakota2edge2stress	Main stress on second and penult, secondary on fourth syllable
h29	Lakota-alt	Main stress on second and alternate succeeding syllables, but not on last
h30	Lakota2stress-alt	Main stress on second and penult, secondary on fourth and alternate succeeding syllables

In languages such as Latvian, French, Maranungku, Weri and Garawa, primary stress is always on a syllable at the edge of the word. In Lakota and Polish, whose learning times are substantially greater than those of the other languages, primary stress is always at a non-edge syllable, except in mono- and di-syllables. (Paiute and Warao are identical with respect to the placement of primary stress to Lakota and Polish, respectively, but are unlearnable.) Thus, *placement of primary stress* seems computationally relevant. In particular, it appears more difficult to learn patterns in which primary stress is assigned at the edges inconsistently.

To test these assumptions, and to determine what features of Paiute and Warao led to their non-learnability, we constructed *hypothetical* stress patterns that exhibit more combinations of the above features than may actually exist. These stress patterns are described in Table 3. We trained networks on these variations in order to evaluate the effects of the various features.

The following factors emerged as determinants of learnability for the range of QI patterns considered:

(I) Inconsistent Primary Stress (IPS): it is computationally expensive to learn the pattern if neither edge receives primary stress except in mono- and di-syllables; this can be regarded as an index of computational complexity that takes the values $\{0, 1\}$: 1 if an edge receives primary stress inconsistently, and 0, otherwise.

(II) Stress clash avoidance (SCA): if the components of a stress pattern can potentially lead to *stress clash*, then the language may either actually permit such stress clash, or it may avoid it. This index takes the values $\{0, 1\}$: 0 if stress clash is permitted, and 1 if stress clash is avoided.

(III) Alternation (Alt): an index of learnability with value 0 if there is no alternation, and value 1 if there is. Alternation means a pattern of some kind that repeats on alternate syllables.

(IV) Multiple Primary Stresses (MPS): has value 0 if there is exactly one primary stress, and value 1 if there is more than one primary stress. It has been assumed that a repeating pattern of primary stresses will be on alternate, rather than adjacent syllables. Thus, [Alternation=0] implies [MPS=0]. The hypothetical stress patterns examined include some with more than one primary stress; however, as far as is known, no actually occurring QI stress pattern has more than one primary stress.

(V) Multiple Stress Levels (MSL): has value 0 if there is a single level of stress (primary stress only), and value 1 otherwise.

It is possible to order these factors with respect to each other to form a five-digit binary string characterizing the ease/difficulty of learning. That is, the computational complexity of learning a stress pattern can be characterized as a 5-bit binary number whose bits represent the five factors above, in decreasing order of significance. Table 2 shows that this characterization captures the learning times of the QI patterns quite accurately. As an example of how to read Table 2, note that Garawa takes longer to learn than Latvian (165 vs. 17 epochs). This is reflected in the parameter setting for Garawa, “01101”, being lexicographically greater than that for Latvian, “00000”.

The analysis of learnability is summarized in Table 4 for all the QI stress patterns, both actual and hypothetical. It can be seen that the 5-bit characterization fits the learning times of various actual and hypothetical patterns reasonably well; there are, however, exceptions, indicating that this 5-bit characterization is only a heuristic. For example, the hypothetical stress patterns with reference numbers h21 through h25 have a higher 5-bit characterization than some other stress patterns, but lower learning times.

The effect of **stress clash avoidance** is seen in consistent learning time differentials between stress patterns of complexity less than or greater than binary “1000”. Learning times with complexity “0001” are in the range 10 to 25 epochs, while complexity “1001” patterns are of the order of 170 epochs; complexity “0010” is of the order of 30 epochs, and “1010”, 190 epochs, for patterns the only difference between which is the absence/presence of stress clash avoidance (Latvian2edge and Latvian2edge-SCA, references h5 and h19 respectively). A pattern with complexity “0011” (Latvian2edge2stress, reference h6) has a learning time of 37 epochs, while a pattern differing only in the addition of SCA (Latvian2edge2stress-SCA, reference h20) takes 206 epochs. Complexity “0101” patterns are in the range 30 to 60 epochs, while complexity “1101” patterns are in the range 70 to 170 epochs; in particular, while Garawa (reference L5) has a learning time of 165 epochs, the same pattern without SCA has a learning time of 38 epochs (Garawa-SC, reference h10). A stress pattern of complexity “0111” takes 85 epochs to learn (Latvian2edge2stress-1alt, reference h16), while addition of stress clash avoidance results in a learning time of 129 epochs (Latvian2edge2stress-1alt-SCA, reference h25).

The effect of **alternation** is seen in a contrast between learning times for patterns of complexity “001” (range: 10 to 25 epochs) and “101” (range: 30 to 60 epochs); “010” and “110” (30 epochs vs. a range of 60 to 90 epochs); “011” and “111” (37 vs. 85 epochs).

The effect of **multiple primary stresses** is seen in the contrast between the stress patterns: Latvian2stress (reference h1, complexity “001”, 21 epochs) and Latvian2edge2stress (reference h6, complexity “011”, 37 epochs); Latvian2edge2stress-alt (reference h9, complexity “101”, 58 epochs) and Latvian2edge2stress-1alt (reference h16, complexity “111”, 85 epochs).

The effect of **inconsistent primary stress** is considerable: stress patterns with the most significant bit 1 are learnable in the perceptron model only if all the other bits are 0; such patterns (Lakota, Polish, references L6, L7, complexity “10000”, 255 epochs) have a higher learning time than any of the patterns with most significant bit 0. All the examined stress patterns of complexity greater than “10000” were unlearnable in the perceptron model. Recall that Paiute and Warao were unlearnable; the present framework is consistent with that result, since under the present analysis, these two patterns have complexity “10101”. Recall also from footnote 8 that if we assume no stressed monosyllables in Paiute and Warao, then these stress patterns become learnable, in approximately 35-45 epochs. Under these circumstances, the complexity index of the two stress sytems is “00101”. As can be seen from Table 2 and Table 4, a learning time of approximately 40 epochs is completely consistent with the learning times obtained for other stress patterns with this index. That is, our scheme for indexing complexity is consistent with the results for Paiute and Warao, whether or not there are assumed to be stressed monosyllables.

The impact of **multiple stress levels** is relatively smaller, and less uniform, both of which motivate this factor’s being treated as least significant. Thus though there are several instances where a stress pattern with a greater number of stress levels has a higher learning time (h1 vs. L1; h3 vs. L2; h6 vs. h5; h7 vs. L3; h8 vs. L4; h21 vs. L5), there

Table 4: Analysis of Quantity-Insensitive learning.

IPS	SCA	Alt	MPS	MSL	LANGUAGE	REF	EPOCHS (syllabic)
		0	0	0	Latvian	*L1*	17
					French	*L2*	16
		0	0	1	Latvian2stress	h1	21
					Latvian3stress	h2	11
					French2stress	h3	23
					French3stress	h4	14
		0	1	0	Latvian2edge	h5	30
		0	1	1	Latvian2edge2stress	h6	37
		1	0	0	*impossible*		
		1	0	1	Maranungku	*L3*	37
					Weri	*L4*	34
					Paiute (unstressed monosyllables)	*L8′*	45
					Warao (unstressed monosyllables)	*L9′*	36
					Maranungku3stress	h7	43
					Weri3stress	h8	41
					Latvian2edge2stress-alt	h9	58
					Garawa-SC	h10	38
					Garawa2stress-SC	h11	50
		1	1	0	Maranungku1stress	h12	61
					Weri1stress	h13	65
					Latvian2edge-alt	h14	78
					Garawa1stress-SC	h15	88
		1	1	1	Latvian2edge2stress-1alt	h16	85
	1	0	0	0	*impossible*		
	1	0	0	1	Garawa-non-alt	h17	164
					Latvian3stress2edge-SCA	h18	163
	1	0	1	0	Latvian2edge-SCA	h19	194
	1	0	1	1	Latvian2edge2stress-SCA	h20	206
	1	1	0	1	Garawa	*L5*	165
					Garawa2stress	h21	71
					Latvian2edge2stress-alt-SCA	h22	91
	1	1	1	0	Garawa1stress	h23	121
					Latvian2edge-alt-SCA	h24	126
	1	1	1	1	Latvian2edge2stress-1alt-SCA	h25	129
1		0	0	0	Lakota	*L6*	255
					Polish	*L7*	254
1		0	0	1	Lakota2stress	h26	**
1		0	1	0	Lakota2edge	h27	**
1		0	1	1	Lakota2edge2stress	h28	**
1		1	0	1	Paiute	*L8*	**
					Warao	*L9*	**
1		1	1	0	Lakota-alt	h29	**
1		1	1	1	Lakota2stress-alt	h30	**

Note. These learning times were obtained using the *syllabic* input representation. IPS=Inconsistent Primary Stress; SCA=Stress Clash Avoidance; Alt=Alternation; MPS=Multiple Primary Stresses; MSL=Multiple Stress Levels. Symbols L1-L9 refer to Table 1, and h1-h30 to Table 3.

Table 5: Summary of results and analysis of QI and QS learning.

Agg	IPS	SCA	Alt	MPS	MSL	QI LANGUAGES	REF	EPOCHS (wt-string)	QS LANGUAGES	REF	EPOCHS (wt-string)
0	0	0	0	0	0	Latvian French	L1 L2	2 2			
0	0	0	0	0	1				Koya	L10	2
0	0	0	0	1	0				Eskimo	L11	3
0	0	0	1	0	1	Maranungku Weri	L3 L4	3 3			
0	0	1	1	0	1	Garawa	L5	7			
0	0.25	0	0	0	0				Malayalam Yapese	L12 L13	19 19
0	0.50	0	0	0	0				Ossetic Rotuman	L14 L15	30 29
0	1	0	0	0	0	Lakota Polish	L6 L7	10 10			
0	1	0	1	0	1	Paiute Warao	L8 L9	** **			
1	0	0	0	0	0				Komi Cheremis	L16 L17	216 212
2	0	0	0	0	0				Mongolian Mayan	L18 L19	2306 2298

Note. These learning times were obtained using *weight-string* input representations. Agg=Aggregative Information; IPS=Inconsistent Primary Stress; SCA=Stress Clash Avoidance; Alt=Alternation; MPS=Multiple Primary Stresses; MSL=Multiple Stress Levels. Symbols L1-L19 refer to Table 1.

are also cases in which a stress pattern with a higher number of stress levels has a lower learning time than one with fewer stress levels (h2 vs. h1; h4 vs. h3; h11 vs. h15; h21 vs. h23).

The effects of a particular factor seem to be reduced when a higher-order bit has a non-zero value. Thus, the effects of alternation are less clear when there is stress clash avoidance: without SCA, the range of learning times for patterns without alternation is 10 to 40 epochs, and with alternation 30 to 90 epochs; but with SCA, the range without alternation is 160 to 210 epochs, and with alternation 70 to 170 epochs.

In summary, the "complexity measure" suggested here appears to identify a number of factors relevant to the learnability of QI stress patterns within a minimal connectionist architecture. It also assesses their relative impacts. The analysis is undoubtedly a simplification, but it provides a framework within which to relate the various learning results.

3.2. Incorporating QS Systems

We now turn to a consideration of quantity-sensitive (QS) stress systems. For QS patterns, information about syllable weight needs to be included in the input representation – the input has to consist of (encoded) sequences of "H" and "L" tokens. A purely *syllabic* input representation is, by definition of quantity-sensitivity, inadequate. *Weight-string* representations were therefore adopted, as discussed in Section 2.2. To maintain consistency of analysis across QI and QS stress patterns, simulations for the QI languages were re-run using the weight-string representation. Note that the stress patterns for all possible weight strings of length *n* are the same for a QI language.

Learning times are shown in Table 5 for simulations of all QI and QS stress systems. The section on the left shows QI learning times, while the section on the right shows QS learning times. Note that in this table, learning times for all languages are reported in terms of the weight-string representation rather than the unweighted syllabic representation used for the previous QI studies.

The differences in learning times across QI patterns are less marked than the differentials in Table 2. This is the result of the increased training set size with the weight-string representation as compared with the syllabic representation. However, learning times are completely consistent with the overall learnability analysis developed in the previous section, so that the analysis of QI systems is in no way affected. The reader may wish to compare the learning times for QI systems in Table 5 (weight-string input representations) with those in Table 2 (syllabic input representations).

In order to incorporate QS systems, the learnability analysis proposed in Section 3.1 on the basis of QI patterns turns out to require some refinement (although the analysis of QI systems themselves will remain unchanged; see below). Inconsistent Primary Stress (IPS) was previously hypothesized as taking binary values; a value of 1 for IPS indicated that primary stress was assigned inconsistently at the edge of words; a value of 0 indicated that this was not the case. If this measure is modified so that its value indicates *the proportion of cases in which primary stress is not assigned at the edge of a word*, the learning results for both QI and QS patterns can be integrated, to a large extent, into a unified account. Recall that learning times for QS systems are those shown in the right-hand section of Table 5.

The learning times for Malayalam and Yapese are approximately 20 epochs, while those for Ossetic and Rotuman are approximately 30 epochs. The difference between these pairs of stress patterns is: for Malayalam and Yapese, primary stress is placed at the edge *except* when the edge vowel is short and the next vowel long (i.e., except 0.25 of the time); for Ossetic and Rotuman, primary stress falls at the edge *except* when the edge vowel is short, i.e., except in 0.5 of the cases.

The five factors discussed earlier were: Inconsistent Primary Stress (IPS); Stress Clash Avoidance (SCA); Alternation (Alt); Multiple Primary Stresses (MPS); and Multiple Stress Levels (MSL). The values of these indices respectively, for both Malayalam and Yapese, are [0.25 0 0 1 0], and for both Ossetic and Rotuman, [0.5 0 0 1 0]. The difference between learning times for these pairs of otherwise identical patterns can then be accounted for in terms of differing values of the IPS measure. Note that the earlier analysis of QI languages remains unchanged: stress patterns that had IPS value 0 still do, and those that had IPS value 1 still do as well.

The learning times of Komi and Cheremis are substantially higher than those of Koya, Eskimo, Malayalam, Yapese, Ossetic and Rotuman. Recall the stress pattern of Komi: stress the first heavy syllable, or the last syllable if there are no heavy syllables. That is, for Komi, a particular syllable S receives primary stress under the following conditions: (1) there are no heavy syllables to the left of S in the syllable string; and (2) S is Heavy *or* S is the last syllable. The second clause of the conditional involves *single-positional* information: information either about the syllable S itself (S is Heavy), or about the absence/presence of a syllable right-adjacent to S in the weight-string. (If there is no syllable to the right of S in the weight-string, then S is the last syllable; if there is a syllable right-adjacent to S, then S is not the last syllable). The first clause of the conditional, however, involves *aggregative* information: information about *all* the syllables to the left of S in the weight-string. We see that in assigning stress to syllable strings in Komi, one must somehow "scan" the input string to extract this aggregative information; similarly for Cheremis.

Komi and Cheremis can therefore be analyzed as stress patterns that require *aggregative* information for the determination of stress placement; none of the other stress patterns require such information. For example, for Koya, a syllable S should receive stress if it is the first syllable (which can be determined from information about the presence/absence of a syllable in the left-adjacent weight-string position), or if it is heavy, both of which are *single-positional* kinds of information. For Ossetic, a syllable S should be stressed if (a) it is the first syllable *and* it is heavy (which requires *single-positional* information about the left-adjacent weight-string position, and about S itself); or if (b) it is the second syllable (*single-positional* information about the weight-string element two positions to the left of S) *and* the syllable in the left-adjacent position is light (also *single-positional* information).

The difference in learning times between Komi and Cheremis on the one hand, and Koya, Eskimo, Malayalam, Yapese, Ossetic and Rotuman, on the other, can now be analyzed in terms of differing informational requirements. Whether or not *aggregative* information is needed therefore seems to be a further factor relevant to the learnability of stress patterns.

We further need to consider the fact that the patterns of Mongolian and Mayan have very much higher learning times than those of any other stress patterns, including Komi and Cheremis. Recall the stress pattern of Mongolian: stress the first heavy syllable, or the first syllable if there are no heavy syllables. Notice how it differs from that of Komi, whose pattern is: stress the first heavy syllable, or the *last* syllable if there are no heavy syllables.

Thus, for Mongolian, if the current syllable S is heavy it should receive stress if it is the *first* heavy syllable. If the current syllable S is light, then it should receive stress only if (a) there is no syllable to its left in the weight-string (which indicates that it is the first syllable), *and* (b) there is no heavy syllable to its right in the weight-string. That is, for Mongolian, there is *aggregative* information required about heavy syllables both to the left of the current syllable *and* to its right. Information about heavy syllables to the left of S is relevant if S is heavy; information about heavy syllables to the right is relevant if S is light. This requirement for keeping track of dual kinds of aggregative information seems to be what makes the pattern so difficult to learn. This contrasts with Komi, for which aggregative information is required only about syllables to the left of S.

A parallel analysis can be made for Mayan. For both Mongolian and Mayan, then, the very high learning times result from the requirement for two kinds of aggregative information, in contrast with only one kind for Komi and Cheremis. Acquiring this information requires bi-directional scanning of the input string for Mongolian and Mayan, as compared with unidirectional scanning, as discussed above, for Komi and Cheremis.

The results from Komi, Cheremis, Mongolian and Mayan thus suggest an additional factor that is relevant for determination of learnability, but that comes into play only in the case of QS patterns: how much scanning is required for *aggregative* information. This can be treated as a sixth index of computational complexity that takes on the values $\{0, 1, 2\}$. We therefore have the following factor, in addition to the five previously discussed:

(VI) Aggregative Information (Agg): has value 0 if no aggregative information is required (*single-positional* information suffices); value 1 if unidirectional scanning is required (Komi, Cheremis); and 2 if bidirectional scanning is required (Mongolian, Mayan).

With these modifications (viz., refinement of the IPS measure, and addition of the Aggregative measure), the same parameter scheme can be used for both the QI and QS language classes, with good learnability predictions *within* each class, as shown in Table 5.

We note, moreover, that learning times for the QS stress patterns of Koya and Eskimo fit right in with the analysis. The characterization of Koya in terms of our complexity measure is "000001", while that of Eskimo is "000010". Learning times for these systems, accordingly, fit in between the learning times for the QI systems of Latvian and French (complexity index "000000") and Maranungku and Weri (complexity index "000101").

As can also be seen, differences in learning times between QS stress patterns also fit in more generally with the analysis developed earlier, and with the analysis of *single-positional* vs. *aggregative* informational requirements developed in this section.

Thus, both the QI and QS results fall into a single analysis within this generalized parameter scheme and weight-string representation, although with a less perfect fit than the within-class results. For example, the QI stress patterns of Lakota and Polish have higher complexity indexes than the QS stress patterns of Malayalam, Yapese, Ossetic and Rotuman, but lower learning times. Quantity-sensitivity thus appears to affect learning times, as seems reasonable to expect, due to the distribution of the weight-string training set (see discussion in Section 2.2). However, no "measure" of its effect will be offered here. The analytical framework developed thus far appears to hold within QI languages, and within QS languages; further analysis would be needed to relate learning results across the two kinds of stress patterns.

Finally, it is worth noting that learning times for these stress patterns have also been obtained with a three-layer architecture. Differences between learning times for different stress patterns were somewhat reduced. However, learning times were *ordered* exactly as in the two-layer model results shown in Table 5, suggesting that these learning results are robust with regard to architectural differences. In the discussion throughout this paper, however, we focus on results from the two-layer model only, since the two-layer model with a single output unit facilitates analysis of connection weights and their contribution to processing.

3.3. Markedness and Learning Times: Correspondences with Metrical Theory

So far, we have used our perceptron learning results to devise an analytical framework for stress. It should be clear how our theory is analogous to conventional linguistic theory: it is an account of stress systems based on observation of the behavior of a perceptron exposed to those systems, in very much the same way that metrical theory is an account of linguistic stress, based on observations of what human beings learn. The stress pattern descriptions given to the perceptron are at the same level of abstraction as those that form the starting point for metrical theory. But since our account is based on learning time rather than surface forms and distributional data, we refer to it as a pseudo-linguistic theory.

In this section, we will discuss the ways in which our learning time results are in good agreement with some of the "markedness" predictions of metrical theory. That is, our results exhibit a correspondence with theoretical predictions, strengthening the analogy between our theory and metrical theory. As we will show, learning results such as these could also provide an additional source of data for choosing between theoretical alternatives.[9]

Within the dominant linguistic tradition, a universal grammar of stress should incorporate a theory of markedness, so as to predict which features of stress systems are at the core of the human language faculty and which are at the periphery. The *distributional* approach to markedness treats as "unmarked" those linguistic forms that occur more frequently in the world's languages. Another approach to markedness is *learnability theory*, which examines the logical process of language acquisition[10]. Thus, for example, Dresher & Kaye take iteration to be the default or unmarked setting for parameter P11, because there is evidence that can cause revision of this default if it turns out to be the incorrect setting: the absence of any secondary stresses serves as a diagnostic that feet are *not* iterative (Dresher and Kaye, 1990, p. 191). If non-iteration were the default, their learning system might not encounter evidence that would enable it to correct this default setting, if it were in fact incorrect. It should be noted that, while this is a representative application of subset theory, the choice of default parameter values depends on the particular learning algorithm employed.

Table 6 shows the stress systems grouped by their theoretical analyses in terms of the parameter scheme discussed in Section 1.4. The last column of the table shows the average learning time in epochs for each group of stress patterns (these are the same learning times shown in Table 5)[11]. As can be seen, there appears to be a fairly systematic differentiation of learning times for groups of stress patterns with different clusters of parameter settings[12].

Learning times appear to be significantly higher for stress systems in groups 5 through 9, which have non-iterative feet, than for those in groups 1 through 4, which either do not have metrical feet at all, or else have iterative feet. This makes the interesting prediction that non-iterative feet are more difficult to learn, and hence marked. This prediction corresponds with both Halle & Vergnaud's Exhaustivity Condition[13], and with the choice of marked and unmarked settings in Dresher & Kaye's parameter scheme (Parameter P11)[14].

Comparison of learning times for group 1 vis-a-vis groups 2, 3 and 4 also suggests that a stress system with only a word-tree (i.e., with no metrical feet) is easier to learn than one with (iterative) metrical feet.

The dramatic difference in learning times between groups 8 and 9 suggests that it is marked for the dominant node to be obligatorily branching[15]. Group 8 differs from group 9 only in not having obligatory branching, and average learning times were 214 epochs vs. 2302 epochs.

---

[9] Nyberg's parameter-based model also provides such data, but in terms of *how many examples* of a stress pattern have to be presented to the stress learner (Nyberg, 1990).

[10] As an example, the Subset Principle (Berwick, 1985; Wexler and Manzini, 1987) has implications for markedness. Suppose that two possible settings a and b for parameter P result in the learner respectively accepting sets S_a and S_b of linguistic forms. If S_a is a subset of S_b, then, once P has been set to value b, no positive evidence can ever re-set it to a, even if that was the correct setting. Unmarked values for parameters should therefore be the ones yielding the most constrained system.

[11] As noted in Section 2.2, *weight-string* representations are necessary for QS stress patterns. For QI systems, *syllabic* representations are sufficient. Of course, weight-string representations can be used for QI systems, although the information about syllable weight will be redundant. To obtain learning times that might reflect differences between the stress patterns themselves, rather than merely reflecting differing input representations and training set sizes, we ran simulations for both QS and QI patterns using weight-string input representations. The learning times in Table 6 are based on use of this weight-string representation for both QI and QS patterns.

[12] The stress system of Garawa, as described in Hayes (1980, p. 54-55), cannot be characterized in terms of the parameter scheme adopted here. The stress systems of Paiute and Warao cannot be learned by a perceptron, as already discussed in Section 3.1, and as will be discussed in more detail in Section 4.3. (Recall, however, that they can be learned (a) using a two-layer architecture with two output units, or (b) a three-layer architecture with "hidden" units, or (c) within the present perceptron architecture, if monosyllabic stress is not included.) Learning results for these three stress systems have therefore been excluded from the present analysis, as either the parameterized characterization or learning times cannot be established.

[13] "The rules of constituent boundary construction apply exhaustively ..." (Halle and Vergnaud, 1987a, p. 15).

[14] In Dresher & Kaye's model, iteration is the default or unmarked parameter setting because there is evidence that can cause revision of this default. The absence of any secondary stresses serves as a diagnostic that feet are *not* iterative (Dresher and Kaye, 1990, p. 191).

[15] This means that the strong branch of a foot must dominate a heavy syllable, and cannot dominate a light one.

Table 6: Learning times for QI and QS stress patterns, grouped by theoretical analysis.

	LANGUAGE	CHARACTERIZATION	EPOCHS
1	Latvian, French	Word-tree, no feet	2
2	Koya	Word-tree, iterative unbounded QS feet	2
3	Eskimo	No word-tree, iterative unbounded QS feet	3
4	Maranungku, Weri	Word-tree, iterative binary QI feet	3
5	Lakota, Polish	Word-tree, non-iterative binary QI feet	10
6	Malayalam, Yapese	Word-tree, non-iterative binary QS feet, dominant node branches	19
7	Ossetic, Rotuman	Word-tree, non-iterative binary QS feet	29 ($\pm$1)
8	Komi, Cheremis	Word-tree, non-iterative unbounded QS feet	214 ($\pm$2)
9	Mongolian, Mayan	Word-tree, non-iterative unbounded QS feet, dominant node branches	2302 ($\pm$4)

Note. These learning times were obtained using *weight-string* input representations for both QI and QS patterns. Each figure is the average learning time for languages in the group.

This prediction agrees with the distributional view that obligatory branching is relatively marked[16], but runs counter to Dresher & Kaye's choice of default values (parameter P7)[17].

However, comparison of group 6 with group 7 suggests that systems with obligatory branching may be more *easily* learned: group 6, with obligatory branching, has a learning time of 19 epochs, compared with group 7, without obligatory branching, but with a learning time of 29 epochs. This runs counter to the distributional argument, but agrees with the learnability view.

Two points are worth noting. First, it is interesting that where there is a conflict between the distributional and learnability theory predictions of markedness, there is also conflicting evidence from the perceptron simulation. Second, these conflicting perceptron results highlight the fact that it may be infeasible to analyze the effects of different settings for *individual* parameters; it may only be possible to make broader analyses of the effects of *clusters* of parameter settings. Strong interactions between parameters have also been observed in other computational learning models of metrical phonology (Eric Nyberg, personal communication).

In view of the greater differential in learning times between Groups 8 and 9 than between Groups 6 and 7, we conclude that the effect of obligatory branching is to *increase* learning time. That is, we view our learning results as supporting the markedness of obligatory branching. This raises the interesting possibility that learning results such as those from the present perceptron simulations can provide a new source of insight into questions of markedness. The distributional view of the markedness of obligatory branching (Hayes, 1980, p. 113) seems to conflict with the learnability view (Dresher and Kaye, 1990, p. 193). The present simulations seem to agree with the distributional view, and could perhaps serve as a fresh source of evidence.

This potential contribution to theoretical analysis can be further illustrated for the stress systems of Lakota and Polish, which are mirror images. Recall that in Lakota, primary stress falls on the second syllable of the word. The analysis so far adopted for Lakota is that it has non-iterative binary right-dominant QI feet constructed from left to right, with a left-dominant word-tree[18]. Let us call this *Analysis A*. As illustrated in Figure 3, this leads to the construction of one binary right-dominant QI foot at the left edge of the word. This, together with the left-dominant word-tree, results in the assignment of primary stress to the second syllable. As has been shown, under this analysis the perceptron learning results support the markedness of non-iteration (recall the differing learning times of Groups

[16]Thus, Hayes (1980, p. 113): "... the maximally unmarked labeling convention is that which makes all dominant nodes strong ... the convention that wins second place is: label dominant nodes as strong if and only if they branch ..."

[17]Obligatory branching is the default because evidence (the presence of any stressed light syllables that do not receive stress from the word-tree) can force its revision (Dresher and Kaye, 1990, p. 193).

[18]This is based on Hayes' analysis of *penultimate* stress (Hayes, 1980, p. 55).

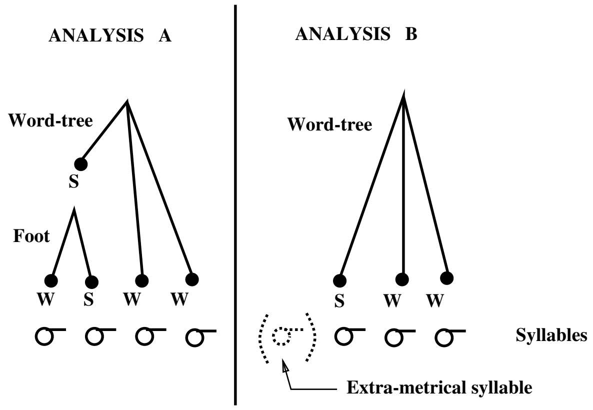

Figure 3: Two metrical analyses for a four-syllable word in Lakota. Strong branches are labeled "S", and weak branches "W". *Analysis A*: construction of one binary right-dominant QI foot at the left edge of the word, together with a left-dominant word-tree, results in the assignment of primary stress to the second syllable. *Analysis B*: the leftmost syllable is treated as "invisible" to the stress rules (*extrametrical*), and the word-tree assigns primary stress to the leftmost of the "visible" syllables. The result is that the second syllable receives primary stress.

1 through 4, vs. Groups 5 through 9).

However, an alternative analysis is that Lakota has a left-dominant word-tree with no metrical feet, and the first syllable is extrametrical (Dresher and Kaye, 1990, p. 143). Let us call this *Analysis B*. As illustrated in Figure 3, the leftmost syllable is treated as "invisible" to the stress rules, and the word-tree assigns primary stress to the leftmost of the "visible" syllables. The result is that the second syllable receives primary stress. Under this analysis, Lakota and Polish (Group 5, in Table 6) differ from Latvian and French (Group 1 in Table 6) only in having an extrametrical syllable. The differing learning times for the two groups (1 epoch vs. 10 epochs) then suggest that extrametricality is *marked*. However, this runs counter to both the distributional view (Hayes, 1980, p. 82)[19] and the learnability theory view (Dresher and Kaye, 1990, p. 189, 191)[20].

To summarize, *Analysis A* views Lakota and Polish as having non-iterative feet, which both the distributional/theoretical and learnability approaches treat as marked. *Analysis B* views these stress patterns as having an extrametrical syllable, which both approaches treat as unmarked. So far, there is nothing theory-external to help choose between the analyses. Simulation results such as these provide such a means: since the learning results are consistent with the theoretical markedness of non-iteration, but not with the unmarkedness of extrametricality, they provide at least weak support for preferring *Analysis A* over *Analysis B*.

4. LOWER-LEVEL PROCESSING: THE CAUSAL ACCOUNT

In the previous section, we developed our abstract "pseudo-linguistic theory" of the learning and assignment of stress by a perceptron. This high-level account was based on observation of perceptron learning results, without knowledge of processing mechanisms within the perceptron.

In this section, we open up the "black box" of the perceptron and examine whether our high-level theory is useful in elucidating the lower-level processing mechanisms.

First, we will examine the connection weights that develop within the perceptron when it has learned particular stress patterns. We will show that these connection weights bear a structural resemblance to the metrical theoretic characterizations of the corresponding stress patterns.

---

[19] Hayes argues for the importance of the device of extrametricality to the theory because of its role in accounting for a variety of stress phenomena. Thus, extrametricality is unmarked in the sense previously referred to: that of being "attested to in a fair variety of languages".

[20] In Dresher & Kaye's formulation, the presence of stress on a leftmost or rightmost syllable can rule out extrametricality. However, there is no positive cue that unambiguously determines the presence of extrametricality. Hence the default value for parameter P8 is that there is extrametricality (P8[Yes]). If the default were P8[No], but this was an incorrect setting, there would be no cue that could lead to detection that this is incorrect.

Second, we will show that, despite these suggestive similarities, there is no analogue between the constructs of metrical theory (or our pseudo-linguistic theory) and actual processing in the perceptron. Thus, despite the fact that the perceptron's mechanisms yield (a) connection weights that look like linguistic constructs, and (b) learning results that give rise to a pseudo-linguistic theory, neither of these high-level accounts is useful in understanding low-level processing.

Finally, we will show that certain non-learnability results can be explained only at the level of perceptron processing, and that this explanation has little in common with the non-learnability analyses of our pseudo-linguistic account.

4.1. Connection Weights and Metrical Theory

Connection weights and perceptron learning. In learning a stress pattern, the perceptron has acquired and encoded in its connection weights its "knowledge" of that pattern. Connection weights for the sixteen stress patterns discussed in Section 3.3 are shown in Figure 4. Each display is a representation of the network as a whole. The large grey shaded rectangle represents the input buffer of the network, organized as two rows of 13 values, corresponding to the rows marked "L" and "H" in Figure 2. The single square protruding from the left is the perceptron's *bias* connection. The perceptron unit itself is represented by the protruding square at the top.

A blob in a particular position denotes a weight on that input connection. White blobs denote positive weights, and black blobs negative weights. The size (area) of the blobs is proportional to the absolute magnitude of the weight. Weights are scaled so that the largest absolute magnitude is depicted in each display as a perfect square; other weights in that display appear as blobs of proportionate size. The scale is shown in the title bar of each display. Thus, for Maranungku, the absolute magnitude of the largest weights is 2.18; these are the large (black) negative weights left of center in the input layer.

We extend the numbering scheme for input connections given in Section 2.3. The central weights are numbered w_{0L} and w_{0H}, corresponding to rows "L" and "H"; they are referred to collectively as w_0. The pair of weights immediately to their left is numbered w_{-1}, and so on. Let us consider some examples of the interpretation of connection weights:

For Latvian, the large negative weights w_{-1} enable detection of the left edge of a word: only the first syllable of a word passing through the input buffer from right to left will be unaffected by these weights when it is the "current input", i.e., in position 0 in the buffer (see Section 2.2 for discussion of processing in the networks). When any non-initial syllable of any word is the current input, there will be some other ("previous") syllable to its left in the buffer. Net input to the perceptron will be negative, since the magnitude of w_{-1} is greater than the magnitude of the positive bias weight. The output will therefore be low, denoting zero stress. The initial syllable of any word, however, will have no syllables to its left in the buffer, and so w_{-1} will have no effect. Net input to the perceptron will therefore be positive (from the bias connection), and so the output will be high, representing primary stress. For French, only the last syllable of a word will escape the effect of the large negative weights w_{+1}, and thus only the last syllable will receive stress. Connection weights for French are the mirror image of those for Latvian, just as the stress patterns themselves are mirror images.

For Weri, the largest weights are w_{+1}; these are large negative weights. Consider the processing of, say, a six-syllable word. When the leftmost syllable is the "current input", and the target output is therefore zero stress, there will be four medium-strength positive weights (w_{+2} and w_{+4}) and four medium-strength negative weights (w_{+3} and w_{+5}), roughly canceling each other out, applying to four of the representations of syllabic elements in the buffer. There is also a pair of large negative weights (w_{+1}). The net input will therefore be negative, resulting in an output of zero stress. When the second syllabic element is the current input, the large negative weights w_{+1} still apply, as do the medium positive weights w_{+2} and w_{+4}. However, the medium negative weights applicable are now only w_{+3}; and so the net input is larger than for the previous syllable, producing an output representing secondary stress. A similar pattern of alternation continues for all the syllables of the word: in each case, there will be either a balance of medium positive and negative weights applicable (resulting in zero stress), or one more pair of positive than negative weights, resulting in secondary stress. The exception is the last syllable: when this is the current input, none of the weights w_{+2} through w_{+5} apply. In addition, there will not be the large negative weights w_{+1} to cancel the positive bias connection. As a result, the net input will be higher for this syllable than for any other, resulting, as desired, in an output representing primary stress. An analogous analysis can be made for Maranungku, whose weights are the mirror image of those for Weri.

For Lakota, if the "current input" is a monosyllable, the bias activation triggers primary stress. However, when the

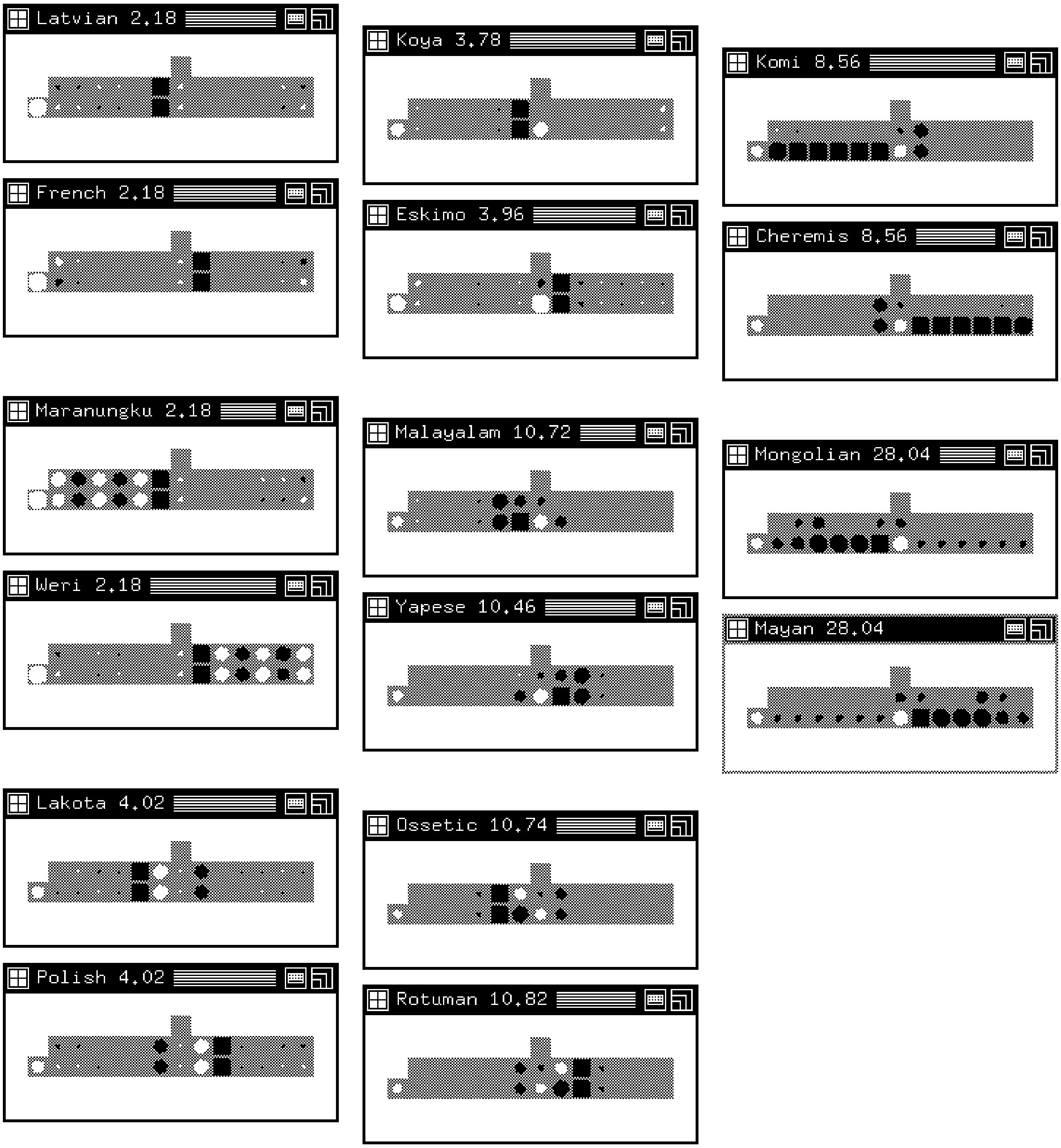

Figure 4: Learned connection weights for sixteen stress patterns. Each display is a representation of the network as a whole. The large grey shaded rectangle represents the input buffer of the network, organized as two rows of 13 values. The single square protruding from the left is the perceptron's *bias* connection. The perceptron unit itself is represented by the single protruding square at the top. A blob in a particular position denotes a weight from the unit in that position to the output unit. White blobs denote positive weights, and black blobs negative weights. The area of the blobs is proportional to the absolute magnitude of the weight. Weights are scaled so that the largest absolute magnitude is depicted in each display as a perfect square; other weights in that display appear as blobs of proportionate size. The scale is shown in the title bar of each display. Thus, for Maranungku, the absolute magnitude of the largest weights is 2.18; these are the large (black) negative weights left of center in the input layer.

current input is the first syllable of a polysyllabic word, the negative weights w_{+1} override the bias activation. If the current input is the second syllable of a word, the perceptron receives high positive activation from w_{-1} in addition to the bias; this is sufficient to overcome the negative weights w_{+1}. However, any syllable after the second triggers the strong inhibitory contribution of w_{-2}, and so cannot receive stress. The analysis for Polish is similar.

The connection weights indicate systematic encoding of knowledge of the patterns by the networks. The fact that two stress patterns are mirror images is reflected in the connection weights as well as in the similarity of learning times.

Correspondences with metrical theory. In this section we will further show that there are correspondences between the form of the encoded knowledge and the characterization of the stress pattern in terms of parameters, thus providing a further analogy between our analyses of perceptron learning and metrical theory. In the discussion below, connection weights for the learned patterns refer to Figure 4. Interpretation of the displays is as discussed above. As in Section 3.3, we will consider the stress systems grouped by their theoretical analyses in terms of Dresher & Kaye's parameter scheme (see Table 6).

Maranungku and Weri are the only stress systems with iterative binary feet (Group 4, Table 6). For these systems, but for no others, there is a very clear binary alternating pattern of positive and negative weights. If, as is natural, we take a positive weight to correspond to the strong branch, and a negative weight to correspond to the weak branch of a foot, then for Maranungku we see left-dominant binary feet, and for Weri right-dominant binary feet – just as in the theoretical analysis[21]. It does not seem too far-fetched to say that the perceptron has discovered a version of iterative binary feet.

The single set of negative weights for Latvian and French (immediately to the left and right of center, respectively) can perhaps be interpreted as a left-dominant and right-dominant word-tree.

Recall that Lakota has non-iterative binary right dominant QI feet constructed from left to right, and that Polish has non-iterative binary left dominant QI feet constructed from right to left. That is, there will be a single binary tree, constructed at the left edge of the word for Lakota, and at the right edge for Polish. Under this analysis, the weights to left and right of center for Lakota and Polish can be interpreted respectively as (single) right-dominant and left-dominant binary QI feet.

The weight patterns for Koya and Eskimo are close to mirror images, but not completely symmetric. Koya assigns primary stress to the first syllable and secondary stress to non-initial heavy syllables, while Eskimo assigns only one level of stress to final and heavy syllables. The chief theoretical difference between the two languages is that the former, but not the latter, has a word-tree. This difference is reflected in the fact that there are two magnitudes, or levels, of connection weights for Koya in the two central buffer positions (the large negative weights in buffer position w_{-1}, and the smaller positive weight in w_0), whereas for Eskimo, there is only one level of weights (the positive and negative weights in w_0 and w_{-1} are approximately equal.) This can be viewed as analogous to the two levels of metrical structure in Koya (metrical feet and word-tree) vs. the single level of structure in Eskimo (metrical feet only.)

Table 6 shows that Malayalam, Yapese, Ossetic and Rotuman (Groups 6 and 7) are the only languages with non-iterative binary QS feet. These are also the only patterns that have more than two large negative weights grouped together to the left (for Malayalam and Ossetic) or right (for Yapese and Rotuman) of center. We can take these three-or-four negative weight structures to correspond to a non-iterative binary QS foot. There is a clear structural difference as compared with the (analogues of) non-iterative binary QI feet in the weights for Lakota and Polish.

Komi, Cheremis, Mongolian and Mayan are the only languages with non-iterative unbounded QS feet (Groups 8 and 9, Table 6). The connection weights for these systems show a pattern of nearly-identical negative weights spanning a set of several input units, and such a pattern does not occur for any of the other stress systems. Such a set of "spanning" weights seems analogous to an unbounded foot. The pattern of weights for Komi seems to correspond to an unbounded right-dominant QS foot, while weights for Cheremis seem to correspond to an unbounded left-dominant QS foot (note the single positive weight at the *right* and *left*, respectively, of the sets of weights, similar to the dominant branch of the foot). The difference in analysis between Komi & Cheremis and Mongolian & Mayan is that feet in the latter pair have obligatory branching, meaning the strong node of the foot must dominate a heavy syllable. As for Komi and Cheremis, the weights for Mongolian and Mayan show a pattern that can be interpreted as an unbounded

[21] As discussed previously, Maranungku has binary, left-dominant QI feet constructed iteratively from the left edge of the word. Weri has binary, right-dominant QI feet constructed iteratively from the right edge of the word.

QS foot. However, they additionally have a set of weights adjacent to the positive weight (i.e., to the "dominant branch" of the unbounded foot), which are not present for Komi and Cheremis; these additional weights can loosely be interpreted as corresponding to a branching dominant node.

4.2. Opening Up the Black Box: Learning Difficulty

It should already be clear from discussion in the previous section that our pseudo-linguistic analysis of Weri as a language of complexity "000101" (see Table 5) does nothing to illuminate the perceptron's processing. Our pseudo-linguistic theory thus provides quite a good predictive framework, but no guidance toward discovery of the perceptron's actual processing mechanisms. In view of the overall resemblance between the nature of our theory and metrical theory, we think the implications are clear: there is no guarantee that linguistic analyses bear a relation to the underlying processing mechanisms, and hence no guarantee that they can usefully guide the search for those mechanisms.

Furthermore, although the connection weights established when the perceptron learns Weri are structurally suggestive of "binary trees", the actual processing involved in the assignment of Weri stress has nothing to do with tree structures. It is interesting that we can "see" binary feet in the connection weights, but this notion does not aid in understanding the perceptron's assignment of Weri stress.

In this section, we will further emphasize these points. As a starting point, recall the earlier analysis of the learning difficulty of Komi, Cheremis, Mongolian and Mayan in Section 3.2. There, we suggested an explanation in terms of types of aggregative information and a "scanning" mechanism. Here, we will examine the learning of these systems in terms of actual processing. We will proceed by analyzing the connection weights for these QS systems.

It may at this point be helpful to review the discussion of perceptron processing in Section 2.3. Note that, with the weight-string representation for inputs, a light syllable is represented by a [1, 0] vector, and a heavy syllable by a [0, 1] vector. For light syllables, therefore, there will be a 1 in the top row (row L) of the input buffer, and a 0 in the bottom row (row H); for heavy syllables, the reverse. Thus, with the weight-string representation the contents of the two rows of the input buffer are usually not identical, and this is relevant to understanding how the connection weights encode knowledge of stress patterns.

Connection weights for the learned patterns are as shown in Figure 4. Interpretation of the displays is as discussed in Section 4.1.

Consider the stress pattern of Koya: primary stress on the first syllable, and secondary stress on heavy syllables. In the weight-diagram for Koya, the bias weight supplies a fairly high positive activation; there is also high positive activation when a heavy syllable is the "current input," arising from w_{0H}. If the current input is the first syllable, then the large negative weights w_{-1} have no effect, and the bias activation results in an output denoting primary stress. If the current input is not the first syllable, then w_{-1} produce a large negative input, whether the syllable in position -1 is heavy or light, thus offsetting positive activation from the bias connection; the net input will be low, resulting in a low output denoting zero stress, *unless* the current input is a heavy syllable, in which case the large positive weight w_{0H} contributes substantially. This positive activation plus that of the bias unit together produce a greater positive net input than is offset by the negative activation from w_{-1}, and so the output is medium, representing secondary stress. In other words, the weights encode the stress pattern: stress the first syllable, and assign secondary stress to heavy syllables.

As a further example, in Malayalam, the current syllable is stressed if it is the first syllable and either it is heavy (large positive activation from w_{0H}, and the large negative weight w_{-1H} has no effect), or it is light but the second syllable is also light (in which case the negative weight w_{+1H} will have no effect). If the current syllable is the first, but is light, and the second syllable is heavy, then w_{0H} will provide no stress, and additionally, w_{+1H} will damp stress provided by the bias connection. If the current input is the second syllable, it receives stress only if it is heavy (positive activation from w_{0H}) and the previous syllable was light (no negative activation from w_{-1H}.) No syllable other than the first or second will be stressed because two of the four large negative weights in w_{-1} and w_{-2} will always be triggered. The analysis for Yapese is similar.

We now turn to an examination of processing and connection weights for Komi and Cheremis. Recall the stress pattern of Komi: stress the first heavy syllable, or the last syllable if there are no heavy syllables. If the current syllable is heavy, it should be assigned primary stress only if there have been no preceding heavy syllables. A heavy current syllable receives stress from w_{0H} and from the bias term, and this is sufficient to offset the effect of the negative weights w_{+1H} and w_{+1L}; but if there is a heavy syllable to its left, this stress is overridden by the weights w_{-1H} through w_{-6H}. Thus a heavy syllable will be stressed *if and only if* it is the first heavy syllable.

If the current input is light, it should be assigned primary stress only if it is the last syllable *and* there have been no heavy syllables in the word. The connection weights make no provision for positive activation from any buffer position containing a light syllable. When a light syllable is the current input, therefore, positive activation comes only from the bias unit; this positive activation, however, is offset by negative activation arising from w_{+1}, and by negative activation arising from w_{-1H} through w_{-6H}. The positive bias is not outweighed by negative activation just in case there are no syllables succeeding the current input in the buffer, and also no heavy syllables preceding the current input, i.e., just in case the current input is the last syllable in a word without any heavy syllables. The analysis of weights for Cheremis is analogous to that for Komi.

Recall that the analysis of Komi in Section 3.2 was that determination of whether or not there are any heavy syllables to the left of the current syllable in the buffer requires *aggregative* information about several syllables; and this suggested the involvement of a special process that "scanned" the buffer.

What is really happening, however, is that the weights w_{-1H} through w_{-6H} determine whether or not there is a heavy syllable preceding the current one. The connection weight displays of Figure 4 illustrate the fact that none of the other QS stress patterns (except Mongolian and Mayan) require establishment of more than two or three weights of large magnitude; for Komi and Cheremis, by contrast, there is a string of large weights across the buffer. It takes longer for the perceptron to learn Komi and Cheremis, therefore, not because it has to learn to perform a special "scanning" operation not required by other languages, but rather, because without weight-sharing it takes a gradient-descent learning algorithm longer to establish a large cluster of weights of significant and roughly equal magnitude, especially when the leftmost buffer elements receive fewer non-zero weight updates than more central elements.[22]

In arriving at this low-level explanation of perceptron learning time differences, our higher-level account was not particularly helpful. Had we been unable to analyze the workings of the perceptron, we might well have hypothesized some lower level mechanism involving a scanning process, and incorporated such into our theory of how the perceptron worked. Alternatively, we might have devised symbolic learning models with an "Aggregation" parameter, whose unmarked value was "no-aggregation-required", and whose marked value was "aggregation-required". In such models, it would naturally take longer to learn a stress system with the the "marked" value than one with the "unmarked" value. The implications for linguistic theory should be clear.

We now again consider the patterns of Mongolian and Mayan, which have very much higher learning times than those of any other stress patterns, including Komi and Cheremis. Compare, once again, the stress patterns of Mongolian and Komi. For Mongolian: stress the first heavy syllable, or the first syllable if there are no heavy syllables. For Komi: stress the first heavy syllable, or the *last* syllable if there are no heavy syllables.

For Mongolian, if the current input is heavy, then it should receive stress if it is the first heavy syllable; thus, as for Komi, each of the weights w_{-1H} through w_{-6H} must be capable of damping the positive activation from w_{0H}. If the current syllable is light, then it should receive stress only if (a) there is no syllable to its left in the buffer (if there is, then one of w_{-1H} or w_{-1L} will override the bias activation), *and* (b) there is no heavy syllable to its right in the buffer. Note that this requires a set of weights w_{+1H} through w_{+6H} to the *right* of the current input, to determine whether there is a heavy syllable.

Thus, for Mongolian, there is a requirement to establish a set of weights to determine the presence of heavy syllables both to the left of the current input *and* to its right, unlike for Komi, which required the establishment of weights only to the left. Furthermore, the weight w_{0H} must be large enough to overcome *all* of w_{+1H} through w_{+6H}, and so must be rather large; but also, each of w_{-1H} through w_{-6H} must be able to override w_{0H}, and so each of these must be even larger. Thus, several very large weights are needed, as evidenced by the magnitude of the largest weights for Mongolian: 28.04, as against a range of approximately 9 to 11 for the other QS patterns. Moreover, there are three levels of weight values to be established, two of which must copied across whole clusters of connections.

Establishing all these weights correctly is what makes the pattern so difficult to acquire using gradient descent learning. The solution space is much further from the initial state due to the large magnitudes of the weights. In addition, there are many more constraints between pairs of non-zero weights that must be satisfied. This is a very different explanation than our high-level account, which hypothesized that the difficulty lay in having to perform dual

[22]We thank Gary Cottrell for pointing out the disparity in error signals seen by different buffer elements.

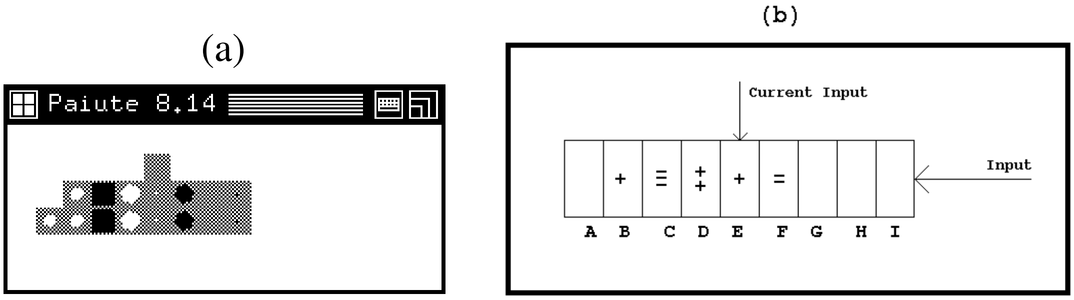

Figure 5: Connection weights for Paiute. (a) Connection weights for the Quantity-Insensitive stress pattern of Paiute, learned for words of up to four syllables. (b) A schematic depiction of those weights in a buffer capable of processing words of up to five syllables. *E* has a modest positive weight to the output unit, and *D* a large positive weight. *F* has a large negative weight, and *C* a very large negative weight.

aggregation by "bi-directional scanning". Again, it should be evident that our pseudo-linguistic account is of little help in determining the nature of actual processing in the perceptron.

4.3. Opening Up the Black Box: Unlearnable Systems

We next re-consider our earlier analysis of the non-learnability of Paiute and Warao. (Recall that in Paiute, primary stress falls on the second syllable, secondary stress on alternate succeeding syllables, and monosyllables were treated as stressed. The pattern for Warao is the mirror image of Paiute.) In Section 3.1, our "parameter" scheme was seen to reflect the non-learnability of these languages in that the 5-bit characterizations of those stress patterns was greater than "10000". We now undertake a more rigorous analysis of those non-learnability results, and show once again that our high-level analysis provides little insight into low-level processing.

The two stress patterns *were* learnable for training sets containing words of up to four syllables (a "length-4" training set); the learning time was 54 epochs. With the addition of five-syllable words, however, (a "length-5" training set), no solution could be found by the perceptron model. The connection weights established for Paiute with the "length-4" training set are displayed in Figure 5a. Figure 5b is a schematic illustration of the same weights. It abstracts away from the weight-string input scheme, which is non-essential for a QI language; the figure also abstracts away the bias connection. The buffer positions *A, B, C, D, E* and *F* in Figure 5b correspond respectively to the positions w_{-4}, w_{-3}, w_{-2}, w_{-1}, w_0 and w_{+1} in Figure 5a. However, 5b depicts a larger buffer than is shown in 5a.

When the first syllable of a four-syllable word is the current input, the large negative weight *F* offsets the positive effect of *E*, resulting in a low net input (and corresponding zero stress). When the current input is the second syllable, the appropriate output is primary stress, which is ensured by the positive activation from *D* and *E* combined, which is greater than the negative activation from *F*. When the third syllable is the current input, the negative weights *C* and *F* combined offset the positive activation of *D* and *E* combined, so that the output corresponds to zero stress. For the fourth syllable, the positive weights *B*, *D*, and *E* combined are sufficiently greater than the negative weight *C* to yield an output corresponding to secondary stress, but not so much greater than *C* as to produce an output corresponding to primary stress.

The processing of the first three syllables of a *five*-syllable word is identical to that just described for a four-syllable word: the weights involved are *C, D, E* and *F*. When the fourth syllable of a *four*-syllable word is the current input, the word is spread over positions *B, C, D* and *E*; the negative weight *F* plays no role. However, when the fourth syllable of a *five*-syllable word is the current input, the five-syllable word is spread over positions *B, C, D, E* and *F*. The output should correspond to secondary stress, just as in the case of the fourth syllable of the four-syllable word, for which the weights *B, C, D* and *E* were appropriate, as described above. For the five-syllable word, however, output is affected not only by those four weights, but also by the weight *F*. For *B, C, D* and *E* to produce the appropriate output in the five-syllable case, *F* would have to be zero; however, for appropriate output on the first syllable of words of length greater than one, *F* must have the negative value shown. Thus there is a conflict in the requirements for *F* for the correct processing of "length-4" and "length-5" training sets.

This is shown more formally below[23] in an argument similar to the now famous proof of non-computability of XOR (Minsky and Papert, 1969). Let θ_1 be the minimum net input required for a response corresponding to primary stress; let θ_2 be the minimum net input required for secondary stress. The following constraints must then hold:

$$
\begin{array}{lll}
(1) & & E > \theta_1 \\
(2) & & F < -E \\
(3) & & E + F < \theta_2 \\
(4) & & \theta_2 > E + F > \theta_1 + F \\
 & \Rightarrow & \theta_2 - \theta_1 > F \\
(5) & & \theta_1 > (B + C + D + E) > \theta_2 \\
 & \Rightarrow & -\theta_1 < -(B + C + D + E) < -\theta_2 \\
(6) & & \theta_1 > (B + C + D + E + F) > \theta_2 \\
 & \Rightarrow & \theta_2 < (B + C + D + E + F) < \theta_1 \\
(7) & & \theta_2 - \theta_1 < F < \theta_1 - \theta_2
\end{array}
$$

Inequality (1) expresses the constraint necessary for E to be able to produce primary stress for a monosyllable. (2) and (3) express the constraints necessary for E and F jointly to be able to suppress both primary and secondary stress for the first syllable of a polysyllabic word. (4) is derived from (3) and (1) as shown, and establishes an upper bound for the magnitude of F. Inequality (5) indicates the constraint that must be satisfied for B, C, D and E to produce secondary stress when the current input is the final syllable of a four-syllable word. So far, the constraints are all as required for correct assignment of stress in the "length-4" training set, and all the inequalities can be satisfied. The additional constraint needed to assign secondary stress to the fourth syllable of a *five*-syllable word is indicated in (6). This is the constraint that makes the "length-5" training set unlearnable. Summing (5) and (6) yields (7), which includes the condition $\theta_2 - \theta_1 < F$. However, we previously have (4) $\theta_2 - \theta_1 > F$. Thus, (4) and (7) impose contradictory constraints on the value of F, as was discussed above. The contradiction is responsible for the non-learnability of Paiute. An analogous demonstration can be made for Warao.

Although Paiute and Warao are in fact learnable by humans (and by three-layer networks, or two-layer networks with two output units), the point here is that any high level description that (for example) declares certain stress patterns unlearnable must have a corresponding low level of processing that grounds it. Our high-level account of the non-learnability of the stress systems of Paiute and Warao was that stress systems with a characterization lexicographically greater than "10000" are unlearnable, and this was supported by the non-learnability of the hypothetical languages h26 through h30. While this analysis is *descriptively* accurate at a high level based on observable properties of these various stress patterns, the *causal explanation* developed in this section can be expressed only in terms of complex low-level interactions of stress pattern with processing architecture. No such explanation is available to ground claims of non-learnability in humans.

There are significant similarities between the development of our six-parameter high-level analysis of perceptron stress learning and the development of the linguistic theory of human stress. Both approaches attempt to identify salient inherent characteristics of the stress systems under examination. Both approaches are concerned with the learnability of the stress systems by the device under study (perceptrons or humans). Both approaches were formulated without reference to the workings of the underlying processor. It is of course not currently possible to open up the black boxes of the human processor. Linguists seem to have assumed, nevertheless, that progress can be made in understanding the nature of those black boxes by using linguistic theory as the guiding framework. In this section, however, we have shown that our pseudo-linguistic theoretical scheme bears little relation to lower-level processing, and has no value in developing a *causal* explanation, because there are no physical correlates of the six parameters. We feel we have provided a persuasive demonstration that linguistic theory may be in much the same position.

---

[23] We thank Geoffrey Hinton for suggesting this approach.

5. THE ROLE OF LINGUISTIC THEORY

In the next two sections we will consider other consequences of the assumption that linguistic analyses can usefully guide the search for processing mechanisms. In Section 5.1 we will argue that the analyses of metrical theory are divorced from real processing also in the sense that their constructs are derived from simplified versions of the data, and involve several levels of abstraction. Processing models based on these analyses therefore suffer some inherent limitations in their ability to account for performance phenomena. Furthermore, there is not even a suggested relationship between these theory-based processing mechanisms and what is known about real neural computation. There seems little reason to assume that the constructs of metrical theory will be appropriate primitives for a processing model.

This, together with some methodological problems discussed in Section 5.2, highlights the need to devise suitable lower-level processing accounts *without* assuming that higher-level theoretical accounts are indicative of what to look for. It is our view that the mechanisms of this lower level will look more like neural network formalisms than like the symbolic structures of the higher level, and that determining the architecture of this lower level will have to be guided by behavioral, neuropsychological, and computational neuroscience investigations. The analyses developed in this paper exemplify how neural network modeling can generate results that *make contact with* the theoretical analyses of the higher level.

5.1. Theoretical Constructs as Processing Primitives

We will examine the data of two stress systems and show that these data have in some cases been significantly simplified in arriving at descriptions such as "stress syllables with branching rimes and closed syllables". Consequently, we will argue, a computational account based on the constructs of metrical theory is necessarily a simplified account. As we noted in Section 1, the models of both Dresher & Kaye (1990) and Nyberg (1992) adopt this default strategy of employing theoretical constructs as primitives. As a result, they are limited in their ability to address processing and performance issues. We will show that, within such a framework, it may not even be possible to *formulate* a processing account of certain stress systems.

West Greenlandic Eskimo. The data on stress in West Greenlandic Eskimo are far less clear than the pattern in Table 1 indicates. Rischel (1974, pp. 91-97) states that the category of stress has no well-defined status in the language's phonology, and that it is very difficult to obtain agreement (from native speakers) on the stress patterns in a variety of word types. There is a strong tendency to hear stress on the last vowel, but Rischel suggests that this may actually be the effect of the intonational contour. Scholars have proposed various stress patterns. Thus Kleinschmidt (1951) suggests that the word has one main accent, and that longer words also have a subsidiary accent which tends to fall either on the first or the last syllable. Very long words may have several subsidiary accents, which are distributed according to the principle that heavy syllables always attract the accent.

Fortescue (1984, p. 340) states that there may be an auditory impression of relative stress on heavy syllables under the influence of certain intonational factors. He suggests that Kleinschmidt's account of stress can probably be reduced to the interaction of syllable weight with intonational nucleus. There may be some residual rhythmicity describable in terms of stress or pitch.

Hayes, citing Schultz-Lorentzen (1945) (which it has not been possible for us to examine), presents a much cleaner description of the stress pattern: stress syllables with branching rimes (i.e., closed syllables) and the final syllable (Hayes, 1980, p. 58). In view of the lack of agreement over Eskimo stress mentioned by both Rischel and Fortescue, this seems to be a significant simplification of the stress pattern.

Koya. In Koya, according to Tyler (1969, pp. 32-33), stress within a word occurs on long syllables, with weak stress on short syllables. Strong stress also occurs on the final syllable under certain circumstances conditioned by the intonation contour. Stress within a phrase occurs on the first syllable.

Hayes' description, adopted in this paper, and citing Tyler, is that primary stress falls on the *first* syllable, and secondary stress on closed syllables or syllables with a long vowel. This corresponds with Tyler's data under the assumption that phrasal stress (falling on the first syllable of the phrase, in Tyler's description) has been conflated with word-level stress. In that case, Tyler's description of stress within a word corresponds with Hayes' description of secondary stress within the word. Intonation-conditioned final-syllable stress also is ignored in Hayes' analysis.

Hayes (1980, p. 58) cites both Koya and West Greenlandic Eskimo as examples of languages that can be analyzed as having unbounded, quantity-sensitive feet, and thus as providing support for the very notion of "unbounded quantity-sensitive foot" and its inclusion in the inventory of metrical theoretic constructs. The adopted descriptions of these stress patterns have been central in the development of metrical theory, but we see that in both cases, the description Hayes adopts is a simplification of the actual stress patterns. Now, to the extent that the details of these stress systems have been simplified, the theory built on these descriptions is an *approximate* theory. That is, metrical theory is a simplified analysis of the phenomena of linguistic stress. While this may appear a trivial or even tautological conclusion, it establishes the first part of our argument that a computational model based on theoretical linguistic constructs is necessarily a simplified high-level model.

Computation and principles and parameters. In the *principles and parameters* approach, an element or rule of linguistic analysis is taken to be part of an innate endowment (a "principle of Universal Grammar") if it is found to be applicable across languages, or to be so abstract that a language learner could not reasonably or logically be expected to learn it from exposure to linguistic data (Hyams, 1986, p. 2). The hypothesis is that the human language faculty is so organized as to make only certain linguistic structures available to human beings. The recurrence of these limited patterns of linguistic structure in the world's languages is taken to be a reflection of the properties of the language faculty.

Language learning is taken to proceed through the discovery of appropriate parameter settings. For this approach to succeed, the relevant linguistic theory must be cast into the "parameter" mold clearly enough to specify (a) what the parameters are taken to be, and (b) what the possible settings of these parameters are, so that a given linguistic system X can be characterized in terms of parameter values P_x. The most explicitly formulated such scheme of which we are aware is that of Dresher & Kaye, discussed in Section 1.4, and adopted here as the theoretical framework.

Now it is clear that there are numerous stress systems that Dresher & Kaye's scheme as presently formulated cannot describe. For example, the stress patterns of Garawa and Aklan do not seem amenable to characterization in terms of these parameters. Hayes' description of Garawa (Hayes, 1980, pp. 54-55) involves three levels of stress, and his analysis involves the construction of binary feet both at the left edge of a word, and iteratively, starting at the right edge of the word. The combination of these operations has no analogue in the parameterized characterization adopted by Dresher & Kaye, whose discussion of Garawa sidesteps this difficulty by simplifying the pattern to just two levels of stress. To take another example, the stress pattern of Aklan is well-known for its complexity; Hayes' analysis (Hayes, 1980, pp. 20-33, page 59) includes conditions that cannot be expressed purely in terms of Dresher & Kaye's parameter scheme, and those authors do not discuss this pattern[24]. Consequently, no existing parameter-based theory can account for these stress systems – which is to say that the parameter-based approach to universal grammar represents a further simplification from the data.

To re-iterate, there are at least two levels of simplification involved. First, the actual data on various stress systems have been simplified to arrive at regular descriptions of those systems (for Koya: "primary stress falls on the initial syllable, secondary stress on closed syllables or syllables with a long vowel" (Hayes, 1980, p. 58)). This is the first level of simplification.

These regular descriptions form the basis of metrical theory, yielding a set of *abstract* constructs in terms of which these stress pattern descriptions can be recast (again for Koya: "feet are unbounded, assigned on the rime projection. Both feet and word trees must be left dominant" (Hayes, 1980, p. 58)).

In a second level of simplification, these abstract metrical constructs have been used to devise a system of parameters. To see that the parameterized formulation involves a further level of simplification from the data, recall that while the stress pattern of Garawa can be characterized in terms of Hayes' metrical theory, it cannot be characterized in terms of the Dresher & Kaye parameter scheme.

It should be clear, therefore, that the account provided by a computational model based on the theoretical parameters of metrical phonology is limited to whatever can be described by the underlying theoretical analysis. First, a

---

[24] We were able to simulate the learning of Aklan using a three-layer version of our model. For a different kind of connectionist treatment of Aklan stress assignment, see Wheeler & Touretzky (1991).

computational account of processing/learning in a particular stress system X can be formulated at all only if the parameterized characterization of X has been formulated. Second, regularities or exceptions in the data that are not captured by the stress pattern as described for the purposes of analysis cannot be dealt with. This concludes our argument that any computational account of learning and/or processing based on principles and parameters will not only be *abstract*, but also *simplified*[25].

Now, it could be argued that a theoretical account is a descriptive formalism, which serves to organize the phenomena by abstracting away from the exceptions in order to reveal an underlying regularity, and that it is therefore a virtue rather than a failing of the theoretical analysis that it ignores "performance" considerations. However, it becomes difficult to maintain this argument with respect to a *processing* model that uses the descriptive formalism as its basis: the processing or learning account still has to deal with actual data and actual performance phenomena.

5.2. Markedness and Impossibility

So far, we have argued that although Universal Grammar is claimed to provide guidance in the search for underlying processing mechanisms, its theoretical constructs may not in fact provide such guidance and constraint.

Now, in linguistic theory, another important motivation for development of a Universal Grammar is the placing of constraints on *possible systems*. Thus Dresher & Kaye argue that one of the motivations for adopting parameters is that they greatly constrain the number of possible stress systems, and rule out crazy non-occurring stress systems that have never been observed (Dresher and Kaye, 1990, pp. 148-151).

The questions of "possibility" and "markedness" are closely related. As we noted in Section 3.3, one approach to markedness is *learnability theory*, which examines the logical process of language acquisition, while the *distributional* approach to markedness treats as unmarked those linguistic forms that occur more frequently in the world's languages.

We will argue in this section that there are methodological problems inherent in both these approaches, in dealing with the theoretically important notions of "markedness" and "possibility". Thus, not only do the theoretical constructs of Universal Grammar not necessarily constrain the search for processing mechanisms, they do not necessarily provide a clear-cut determination of "possible" and "impossible".

Difficulty of determining markedness. Within learnability theory, various proposals have been made regarding the "markedness" of particular grammars. One proposed metric is the number of intermediate grammars that have to be gone through in getting from an initial grammar G_0 to a descriptively adequate grammar G_L for language L (Rouveret and Vergnaud, 1980). That is, the length of the sequence G_0, ..., G_L is a metric of its markedness. Similarly, Williams suggests taking the child's initial hypothesis about the language to be the "unmarked" case (Williams, 1981).

Under this view, the number of times the initial parameter settings have to be revised in arriving at the final parameter settings would indicate a language's complexity or markedness. However, as has been noted previously, the choice of initial or unmarked settings is related to the learning algorithm employed, and the nature of linguistic evidence assumed to be available in a particular model.

As an example of this, consider the notion of extrametricality. In Dresher & Kaye's formulation, the presence or absence of extrametricality is represented by parameter P8 (see Section 1.4). Dresher & Kaye implicitly take the default value for parameter P8 to be P8[Yes], meaning that there is extrametricality (Dresher and Kaye, 1990, p. 189,191). This is because the presence of stress on a leftmost or rightmost syllable can rule out extrametricality; however, there is no positive cue that unambiguously determines the presence of extrametricality. If the default were P8[No], but this was an incorrect setting, there would be no cue that could lead to detection that this is incorrect. In contrast, in Nyberg's model (Nyberg, 1990; Nyberg, 1992), the default value of the same extrametricality parameter is taken to be P8[No], and the performance of his stress learning system indicates that the presence of extrametricality

[25] Our perceptron model corresponds to the first level of simplification noted above: inputs are based on the simplified descriptions. However, no further simplifications are made from the data. Note particularly that the simplifications incorporated in its inputs do not enter into its processing, since these mechanisms do not employ processing primitives *derived* from the data. As a result, while the connectionist model presented in this paper would produce different learning time results if presented with more realistic data, it is not at all clear how a learning system based on the parameters of metrical phonology would perform. For example, there may be no concise, theoretically satisfying parameter set that can describe the actual data of Eskimo, in which case it is hard to see how a parameter-based system could learn that data.

is harder to learn, or marked. Thus, we have two models, based on the same set of parameters[26], one showing that extrametricality is unmarked, and the other demonstrating that extrametricality is marked. Clearly, what is marked or unmarked is by no means an absolute, even within a parameter-based formulation.

It is therefore unclear how much this approach can contribute to determining the "markedness" of different linguistic systems. Obviously, it would be of interest to examine stages of development that children might go through in arriving at the stress pattern of their language. Work by Gerken suggests that children have "different" metrical feet (Gerken, 1991). However, as Dresher & Kaye note (Dresher and Kaye, 1990, p. 42), there seems to be little or no data in the stress acquisition literature relevant to stages of development.

Impossible stress systems. The *distributional* approach to markedness treats as "unmarked" those linguistic forms that occur more frequently in the world's languages. This seems to be the approach taken by, for example, Hayes (1980, p. 50):

> In justifying a foot inventory as the unmarked one, a minimal requirement is to show that all the members of the inventory are attested in a fair number of languages ...

Such an approach can be criticized, however, on the grounds that the frequency of occurrence of some linguistic form does not necessarily determine its status as "core" or "peripheral", and the non-occurrence of some form does not show that it is "impossible." The distribution of languages in the world is a function of many historical, non-linguistic, factors, and does not necessarily have linguistic-theoretic significance. To quote Pullum (1982, p. 343; p. 340):

> ... no one has any idea to what extent the history of the human race has skewed the distribution of [linguistic] types by skewing the distribution of people ... to postulate a default assumption that, say, *wh*-movement cannot be rightward, merely because it is commoner (in currently well-studied languages) for it to be leftward, is surely perverse as well as unnecessary. Language acquisition takes place within the infant, not within the context of a statistical survey of currently attested languages ...

As noted previously, the only linguistic options that can be entertained by the human mind are taken to be those consistent with the principles and parameters of Universal Grammar. Stress systems not sanctioned by the principles and parameters of metrical theory are therefore supposedly impossible. Thus Dresher & Kaye (1990, pp. 148-151) argue that one of the motivations for adopting parameters is that they greatly constrain the number of possible stress systems, and rule out crazy, non-occurring stress systems that have never been observed. One of the criticisms frequently made of connectionist models of language processing is that they can "learn anything", and, in particular, can learn systems not sanctioned by linguistic theory (Pinker and Prince, 1988).

The trouble is that, if this reasoning is carried through, and given Dresher & Kaye's parameters, the stress systems of Garawa and Aklan discussed above are impossible. Of course, it may be possible to extend the parameter scheme so as to describe these systems. But now, "possible" and "impossible" have been reduced to what has or has not been observed in the world. And, as discussed above, such distributional grounding must be viewed with caution. The point we wish to make here, then, is that the principles and parameters notion of "possibility" should be seen as having *heuristic* value rather than as providing definitive prescriptions of what is possible or impossible.

Certainly, the present perceptron model is unlikely to reflect the way that humans learn language, as it would probably be capable of learning "outrageous" patterns quite easily[27]. However, it is important to recognize that judging such a system to be plausible or implausible on these grounds is really an appeal to distributional evidence and intuition. A system of parameters does not provide the litmus of possibility that is sometimes claimed for it.

The only way to settle such questions would seem to be the ability or inability of the child to learn a particular stress system, irrespective of whether or not such a pattern is observed in the world's languages. However, as in the case of determining the *difficulty* of learning, there are no relevant data.

---

[26] Nyberg adopts the Dresher & Kaye scheme as well.

[27] Note that this is a comment on its specific architecture, and not on the general kinds of computational mechanisms it incorporates.

5.3. Recapitulation

The purpose of this paper was to examine an assumption about the role of linguistic theory exemplified in the following quotation (Chomsky, 1988, p.6-8):

> . . . Insofar as the linguist can provide answers [to questions about the form and acquisition of linguistic knowledge], the brain scientist can begin to explore the physical mechanisms that exhibit the properties revealed in the linguist's abstract theory . . . the discoveries of the linguist-psychologist set the stage for further inquiry into brain mechanisms . . .

In Section 3 of the paper, we used observations of the learning behavior of our perceptron model to develop a "pseudo-linguistic theory" of how the perceptron learns various stress systems. This theory is an account of stress systems based on observation of the learning behavior of a perceptron exposed to those stress systems, in very much the same way that metrical theory is an account of linguistic stress, based on observation of the stress patterns that occur in human languages. We showed that our learning results and predictions have correspondences with those of metrical theory. We further showed structural correspondences between the knowledge acquired by our perceptron model when learning various stress patterns and the metrical theoretic characterizations of those stress patterns.

In Section 4, however, we showed that, despite the strikingly suggestive parallels between our theory, perceptron connection weights, and metrical theory, neither our theory nor metrical theory provides any insight into the lower-level processing mechanisms of the perceptron. The way the perceptron computes stress turns out to be completely unrelated to the constructs used in our high-level theory. This is exemplified by the fact that our high-level theoretical account of why Paiute and Warao were unlearnable in the perceptron model bears no relation to the causal explanation of their non-learnability, which could be stated only in terms of low-level processing; and by the fact that our high-level explanations of learning difficulty bear no relation to the actual processes involved in the establishment of connection weights.

In the search for the mechanisms of human language processing, linguistic theory may be as misleading as our pseudo-linguistic theory of the perceptron. The claim that ". . . the discoveries of the linguist-psychologist set the stage for further inquiry into brain mechanisms . . . " is an assumption that may be unwarranted.

Acknowledgements

We would like to thank Deirdre Wheeler for introducing us to the literature on stress and providing guidance in the early stages of this work. We also thank Gary Cottrell, Elan Dresher, Jeff Elman, Dan Everett, Michael Gasser, Brian MacWhinney, Jay McClelland, Eric Nyberg, Brad Pritchett, Steve Small, Deirdre Wheeler, and an anonymous reviewer for helpful comments on earlier versions of this paper, and David Evans for access to computing facilities at Carnegie Mellon's Laboratory for Computational Linguistics. Of course, none of them is responsible for any errors in this work, nor do they necessarily agree with the opinions expressed here.

The second author was supported by a grant from Hughes Aircraft Corporation, and by the Office of Naval Research under contract number N00014-86-K-0678.

References

Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.

Chomsky, N. (1988). *Language and Problems of Knowledge*. MIT Press, Cambridge, MA.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper and Row, New York.

Davis, S. (1988). *Topics in Syllable Geometry*. Garland, New York.

Dresher, B. and Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition*, 34:137–195.

Everett, D. and Everett, K. (1984). On the relevance of syllable onsets to stress placement. *Linguistic Inquiry*, 15:705–711.

Fortescue, M. (1984). *West Greenlandic*. Croom Helm, Beckenham, England.

Gerken, L. (1991). Do adults and children have different feet? In Deaton, K., Noske, M., and Ziolkowski, M., editors, *CLS 26-II: Papers from the Parasession on The Syllable in Phonetics and Phonology, 1990*. Chicago Linguistic Society.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Basil Blackwell, Oxford, England.

Halle, M. and Clements, G. (1983). *Problem Book in Phonology*. MIT Press, Cambridge, MA.

Halle, M. and Vergnaud, J.-R. (1987a). *An Essay on Stress*. MIT Press, Cambridge, MA.

Halle, M. and Vergnaud, J.-R. (1987b). Stress and the cycle. *Linguistic Inquiry*, 18:45–84.

Hayes, B. (1980). *A Metrical Theory of Stress Rules*. PhD thesis, Massachusetts Institute of Technology. Circulated by the Indiana University Linguistics Club, 1981.

Hayes, B. (1984a). Iambic and trochaic rhythm in stress rules. In *Proceedings of the Berkeley Linguistics Society 11*.

Hayes, B. (1984b). The phonology of rhythm in English. *Linguistic Inquiry*, 14:33–74.

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*. Reidel, Dordrecht, The Netherlands.

Kaye, J. (1989). *Phonology: A Cognitive View*. Lawrence Erlbaum, Hillsdale, NJ.

Kleinschmidt, S. (1851). Grammatik der Gronlandischen sprache. IDC Micro-Edition N-282.

Liberman, M. (1975). *The Intonational System of English*. PhD thesis, Massachusetts Institute of Technology. Circulated by the Indiana University Linguistics Club, 1978.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249–336.

McCarthy, J. (1979a). *Formal Problems in Semitic Phonology and Morphology*. PhD thesis, Massachusetts Institute of Technology.

McCarthy, J. (1979b). On stress and syllabification. *Linguistic Inquiry*, 10:443–466.

Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Nyberg, E. (1990). A limited non-deterministic parameter-setting model. In *Proceedings of the North East Linguistics Society*, Université de Québec à Montréal.

Nyberg, E. (1992). *Weight Propagation and Parameter Setting*. PhD thesis, Department of Philosophy, Carnegie Mellon University.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Prince, A. (1976). Applying stress. Unpublished manuscript, University of Massachusetts, Amherst.

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

Pullum, G. (1982). Letter. *Linguistics*, 20:339–344.

Pylyshyn, Z. W. (1973). The role of competence theories in cognitive psychology. *Journal of Psycholinguistic Research*, 2:21–50.

Rischel, J. (1974). *Topics in West Greenlandic Phonology*. Akademisk Forlag, Copenhagen.

Rouveret, A. and Vergnaud, J.-R. (1980). Specifying reference to subject. *Linguistic Inquiry*, 11:97–202.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing*, volume 1: Foundations. MIT Press, Cambridge, MA.

Rumelhart, D. E. and McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processes? In MacWhinney, B., editor, *Mechanisms of Language Acquisition*. Lawrence Erlbaum, Hillsdale, NJ.

Schultz-Lorentzen (1945). *A Grammar of the West Greenlandic Language*. C. A. Reitzel, Copenhagen.

Selkirk, E. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry*, 11:563–605.

Selkirk, E. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Trager, G. and Smith, H. (1951). *An Outline of English Structure*. American Council of Learned Societies, Washington.

Tyler, S. (1969). *Koya: An Outline Grammar*, volume 54 of *University of California Publications in Linguistics*. University of California Press, Berkeley & Los Angeles, CA.

van der Hulst, H. and Smith, N. (1982). An overview of autosegmental and metrical phonology. In van der Hulst, H. and Smith, N., editors, *The structure of phonological representations*, volume I. Foris Publications, Dordrecht, The Netherlands,.

Vergnaud, J.-R. and Halle, M. (1978). Metrical structures in phonology. Unpublished manuscript, Massachusetts Institute of Technology.

Wexler, K. and Manzini, M. (1987). Parameters and learnability in binding theory. In Roeper, T. and Williams, E., editors, *Parameters and Linguistic Theory*. Reidel, Dordrecht, The Netherlands.

Wheeler, D. and Touretzky, D. (1991). From syllables to stress: A cognitively plausible model. In Deaton, K., Noske, M., and Ziolkowski, M., editors, *CLS 26-II: Papers from the Parasession on The Syllable in Phonetics and Phonology, 1990*. Chicago Linguistic Society.

Widrow, G. and Hoff, M. (1960). Adaptive switching circuits. In *Institute of Radio Engineers, Western Electric Show and Convention, Convention Record, Part 4*, pages 96–104.

Williams, E. (1981). Language acquisition, markedness and phrase structure. In Tavakolian, S., editor, *Language Acquisition and Linguistic Theory*. MIT Press, Cambridge, MA.