Cognition, 28 (1988) 195–247

The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models*

JOEL LACHTER
THOMAS G. BEVER
University of Rochester

There's no safety in numbers ... or anything else
(Thurber)

Abstract

Recently proposed connectionist models of acquired linguistic behaviors have linguistic rule-based representations built in. Similar connectionist models of language acquisition have arbitrary devices and architectures which make them mimic the effect of rules. Connectionist models in general are not well-suited to account for the acquisition of structural knowledge, and require predetermined structures even to simulate basic linguistic facts. Such models are more appropriate for describing the formation of complex associations between structures which are independently represented. This makes connectionist models potentially important tools in studying the relations between frequent behaviors and the structures underlying knowledge and representations. At the very least, such models may offer computationally powerful ways of demonstrating the limits of associationistic descriptions of behavior.

1. Rules and models

This paper considers the status of current proposals that connectionist systems of cognitive modelling can account for rule-governed behavior without directly representing the corresponding rules (Hinton & Anderson, 1981; Hanson

*We are grateful for comments on earlier drafts of this paper from Gary Dell, Jeff Elman, Jerry Feldman, Jerry Fodor, Lou Ann Gerken, Steve Hanson, George Lakoff, Steve Pinker, Alan Prince, Zenon Pylyshyn, Patrice Simard, Paul Smolensky, and Ginny Valian. Requests for reprints should be addressed to J. Lachter or T.G. Bever, Department of Psychology, University of Rochester, Rochester, NY 14627, U.S.A. This work was completed while the first author was supported by a National Science Foundation pre-doctoral fellowship.

& Kegl, 1987a, b; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986; Smolensky, in press). We find that those models which seem to exhibit regularities defined by structural rules and constraints, do so only because of their ad hoc representations and architectures which are manifestly motivated by such rules. We conclude that, at best, connectionist models may contribute to our understanding of complex associations between independently defined structures.

For the purposes of our discussion, a rule is a function which maps one representation onto another. The status of rules within the field of cognitive modelling are at the center of a current war. For the past 20 years, the dominant approach to algorithmic modelling of intelligent behavior has been in terms of 'production systems' (for discussions and references see Anderson, 1983; Neches, Langley, & Klar, 1987). Production systems characteristically (but not criterially) utilize a set of statements and algorithmic steps which result in a behavior. Such models characteristically (but not criterially) operate linearly; that is, a model first consults a proposition, applies it if relevant, then goes on to the next proposition, and so on.

Recently, a different paradigm in cognitive modelling has been proposed, using information arranged in systems which can apply as a set of parallel constraint satisfactions. In these systems, a network of interconnected nodes represents the organization underlying the behavior. The relationship between each pair of nodes is an activation function which specifies the strength with which one node's activation level effects another. Such systems are touted as meeting many potential objections to production systems, in that the effect of the nodes on output behavior can be simultaneous, and the relationship to neuronal nets is transparent and enticing (Feldman & Ballard, 1982). Since the nodes are by definition interconnected, this paradigm for artificial intelligence has become known as 'connectionism' (Dell, 1986; Feldman & Ballard, 1982; Grossberg, 1987; Hinton & Sejnowski, 1983; Hopfield, 1982; McClelland & Rumelhart, 1981; for general references on connectionism, see Feldman et al., 1985; McClelland & Rumelhart, 1986).

Connectionist modelling defines sets of computational languages based on network structures and activation functions. In certain configurations, such languages can map any Boolean input/output function. Thus, connectionism is no more a psychological theory than is Boolean algebra. Its value for psychological theory in general can be assessed only in specific models. Language offers one of the most complex challenges to any theoretical paradigm. Accordingly, we concentrate our attention on some recent connectionist models devoted to the description of language behaviors. After some initial consideration of models of acquired language behavior, we turn to models which purport to learn language behaviors from analogues to normal input. Such

models are most important because they might be ambitiously taken as solving the problem of how rule-governable behavior is induced from the environment without the invocation of rules.

What we shall demonstrate, roughly, is this: particular connectionist models for language indeed do not contain algorithmic rules of the sort used in production systems. Some of these models, however, work at their pre-defined tasks because they have special representations built in, which are connectionist-style implementations of rule-based descriptions of language (Dell, 1986; Hanson & Kegl, 1987; McClelland & Elman, 1986). Another model, touted because it does not contain rules, actually contains arbitrary computational devices which implement fragments of rule-based representations—we show that it is just these devices that crucially enable this model to simulate some properties of rule-based regularities in the input data (Rumelhart & McClelland, 1986b). Thus, none of these connectionist models succeed in explaining language behavior without containing linguistic representations, which in speakers are the result of linguistic rules.

2. Models of acquired language behaviors

We consider first some models of adult skill. The goal in these cases, is to describe a regular input/output behavior in a connectionist configuration. Dell's model of speech production serves as a case-in-point (based on Dell, 1986; personal communication; see Figure 1). There are two intersecting systems of interconnected nodes, one for linguistic representations and one for sequencing the output. In the linguistic representation, nodes are organized in four separate levels, words, syllables, phonemes, and phonological features. Each word describes a hierarchy specifying the order of component syllables, which in turn specify their component phonemes, which in turn specify bundles of phonetic features. The sequencing system activates elements in phonologically allowable orders. Each phone receives activation input both from the linguistic subsystem and the sequencing subsystem, which results in their being produced in a specified order. As each phoneme is activated, it in turn activates all the features and all the syllables to which it is connected, even irrelevant ones which are not in the current word; then those features, words and syllables can in turn activate other phonemes. This pattern of radiating activation automatically activates relatively strongly just those irrelevant words and syllables with structurally similar descriptions. Accordingly, the model predicts that errors can occur as a function of the activation of irrelevant phones, syllables and words, but primarily those in structurally similar positions. It is well known that these are just the kind of

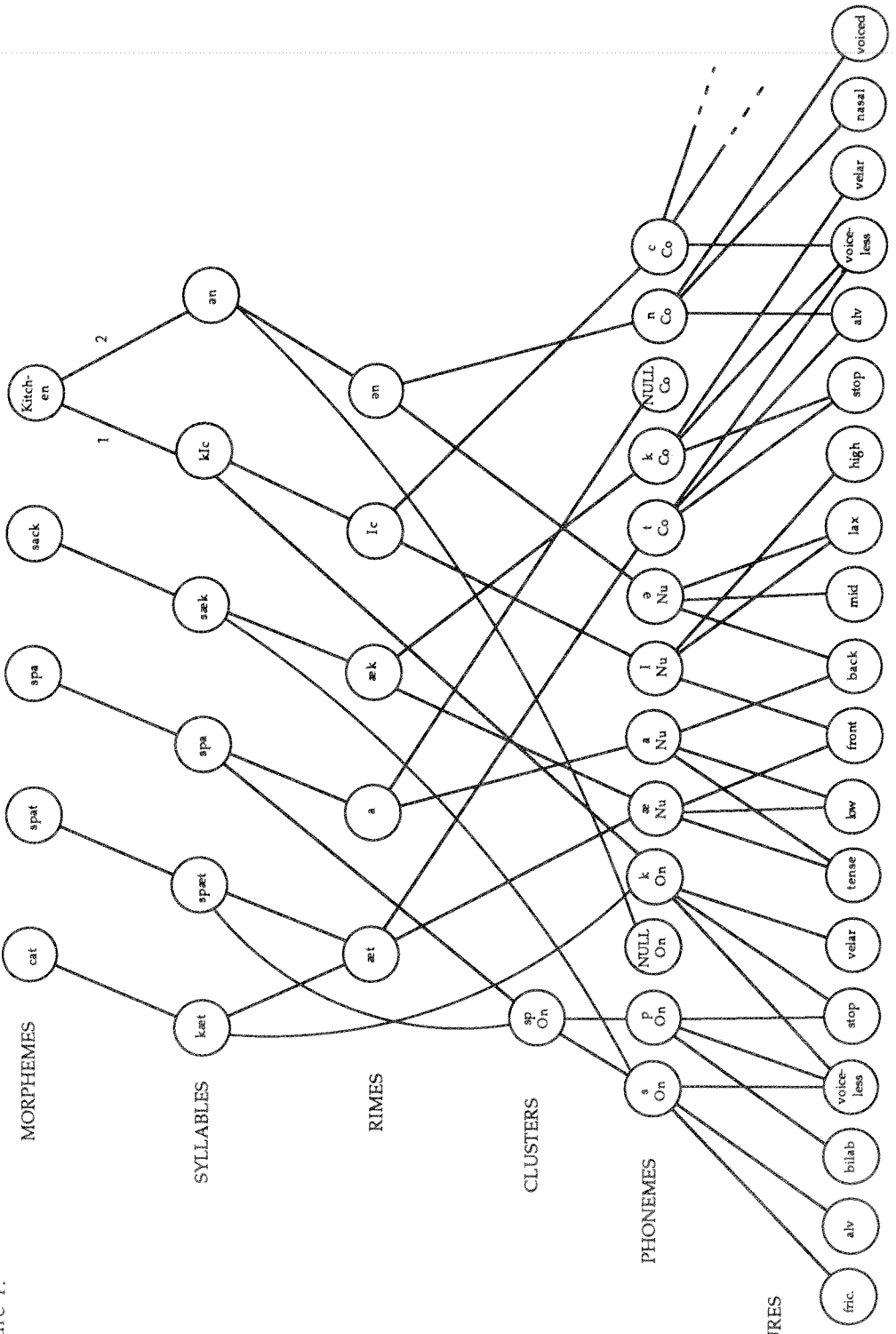

MORPHEMES
SYLLABLES
RIMES
CLUSTERS
PHONEMES
FEATURES
cat
spat
spa
sack
Kitch-
en
1
2
kæt
spæt
spa
sæk
kIc
æt
æk
Ic
an
sp
On
k
On
NULL
On
p
On
s
On
æ
Nu
a
Nu
I
Nu
ə
Nu
t
Co
k
Co
NULL
Co
n
Co
c
Co
fric.
alv
bilab
voice-
less
stop
velar
tense
low
front
back
mid
lax
high
stop
alv
voice-
less
velar
nasal
voiced

Figure 1.

speech errors that occur, exchanges between sounds and words in structurally similar positions.

It is crucial that the nodes are arrayed at different levels of lexical and phonological representation. Each of these levels has some pre-theoretic intuitive basis, but actually the units at each level are consistently defined in terms of a theory with rules which range categorically over those units. Hence, if the model is taken as a psychological theory, it would support the validity of particular levels of representation which are uniquely defined in terms of a rule-governed structural theory. The model also makes certain assumptions about the prior availability of sequence instructions, absolute speed of activation, and speed of inhibition of just-uttered units. Thus, the model comprises a fairly complete outline of the sequencing of speech output, given prior linguistically defined information, and the prior specification of a number of performance mechanisms and parameters. The model is a talking mechanism for the normal flow of speech which can correctly represent errors because it has the linguistic units and the relations between them wired in.

Such models can represent perceptual as well as productive processes. For example, Elman, McClelland and Rumelhart have developed a series of models for word recognition (Elman & McClelland, 1986; McClelland & Elman, 1986; McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). These models are similar to Dell's model in the sense that they have several levels of linguistic representation built into them. For example, TRACE (McClelland & Elman, 1986) recognizes words based on a stylized acoustic feature sequential input. The model utilizes three kinds of detector nodes, for acoustic features, phonemes and words. Feature nodes are grouped into 11 sets, each corresponding to a successive 'time slice' of the acoustic input. Each slice presents the value of each of seven feature dimensions; each dimension has 9 possible values. Accordingly, at the input level of this model, each 'phoneme' is represented in terms of a feature/value/time-slice matrix which defines 693 separate nodes. The centers of each set of time slices from a phoneme are spaced 6 slices apart, to simulate the acoustic overlap that real phones have in speech. Every three time slices, there is centered a set of connections to 15 phoneme detector nodes, with each phoneme node receiving input from 11 time slices. This means that a three-phoneme word activates 1993 feature/value/time-slice nodes and 165 phoneme units. Finally, there are 211 word nodes appropriately linked to their constituent phonemes.

TRACE builds in several levels of linguistic analysis. This is not merely an external observation about the model, it is reflected in its internal architecture. All nodes at the same level of representation inhibit each other, while nodes at adjacent levels of representation excite each other. The within-level inhibition serves the function of reducing the active nodes at each level to

those that are 'most' active; the across-level facilitation serves the function of increasing the effect of one level on another. The within-level inhibition in TRACE is set at distinct values for each level, as are the levels of excitation between different levels. The result is that the model exhibits qualitatively distinct sets of nodes, grouped and layered in the same way as the corresponding linguistic levels of representation. With all this built-in linguistic apparatus, TRACE can model a number of interesting aspects of lexical access by humans. First, it can recognize words from mock acoustic feature input. Second, it can generate a variety of effects involving interference relations between levels of representation: for example, a family of 'word superiority' effects, in which properties of words influence the perception of particular phonemes and features. Such effects occur because nodes at the different levels of representation are interconnected. We take this to be one of the obvious features of connectionist algorithms: insofar as there are interactions between levels of representation—especially influence from higher to lower levels—a parallel system with multiple simultaneous connections both within and between levels is indicated. Indeed, it is difficult to see how a production-system set of rules could be elegantly configured to capture such phenomena.

3. A model which seems to learn past tense rules

The previous models embed an already established linguistic analysis into a model of adult behavior: these models demonstrate that the connectionist framework allows for the implementation of linguistic structure in computationally effective ways. If we take such models as psychologically instructive, they inform us that once one knows what the internal structure of a language is, it can be encoded in such a system in a way which predicts linguistic behavior.

A more ambitious goal is to construct a connectionist model which actually *learns* from regularities in normal input, and generates behavior which conforms to the real behavior. In this way, the model can aspire to explain the patterns characteristic of the behavior. Consider the recent model proposed by Rumelhart and McClelland (1986b, R&M) which seems to learn the past tense rules of English verbs without learning any explicit rules. This model has been given considerable attention in a current review article, especially with reference to its empirical and principled failings and how such failings follow from the choice of computational architecture (Pinker & Prince, 1988). Our approach is complementary: we examine the R&M model internally, in order to understand *why* it works. We argue that the model seems to work for two kinds of reasons; first, it contains arbitrary devices which make it

relatively sensitive to those phonological structures which are involved in the past-tense rules; second, it stages the input data and internal learning functions so that it simulates child-like learning behavior. We first describe the past-tense phenomena in rule-governed terms, and then examine how the R&M model conforms to the consequent regularities in the child's stages of acquisition.

The principles of the past tense formation in English involve a regular ending ('ed') for most verbs, and a set of internal vowel changes for a minority of verbs. In our discussion we are interested in a particular kind of structural rule, typical of linguistic phenomena: one which performs categorical operations on categorical symbols. In this sense, linguistic rules admit of no exceptions. That is, they are not probabilistically sensitive to inputs, and their effects are not probabilistic. We can see how they work in the formation of the regular past tense from the present in English verbs. The regularities are:

mat ... mattED need ... needED	If the verb ends in "t" or "d", add "ed"
push ... pushT pick ... pickT	If the verb ends in sounds, "sh", "k", "p", "ch", ... add "t"
buzz ... buzzD bug ... bugD ski ... skiD	If the verb ends in "z", "g", "b", "j" or a vowel ... add "d"

These facts are described with rules which draw on and define a set of phonological 'distinctive features', attributes which co-occur in each distinct phoneme. Features are both concrete and abstract: they are concrete in the sense that they define the phonetic/acoustic content of speech; they are abstract in the sense that they are the objects of phonological rules, which in turn can define levels of representation which are not pronounced or pronounceable in a given language. The description of the rules of past tense formation exemplify these characteristics. (Note that any particular description is theory dependent. We have chosen a fairly neutral form (basically that in Halle, 1962), although we recognize that phonological theory is continually in convulsions. The generalizations we draw are worded to hold across a wide range of such theories.)

1a. Add a phoneme which has all the features in common to T and D, with the voicing dimension unspecified.
1b. Insert a neutral vowel, 'e', between two word-final consonants that have identical features, ignoring voicing.
1c. Assimilate the voicing of the T/D to that of the preceding segment.

These rules define distinct internal representations of each past form:

	Rule
Push	
PushT/D	1a
PushT	1b
Pit	
PitT/D	1a
PiteT/D	1b
PiteD	1c

There are models in which such rules are interpreted as simultaneous distinct constraints (Goldsmith, 1976) or as applying in strict order (Halle, 1962). Either way, the effects on the output form are categorical, each rule involving a distinct, but abstract phonetic form. The specific shape of the rules also clarifies the sense in which distinctive features are symbolic. For example, rule (1b) refers to sequences of consonants which are made in the same manner and place, regardless of their voicing status. This rule also applies to break up sequences of 's, z, sh, ch' created by affixing the plural and present tense S/Z to nouns ('glitches') and verbs ('botches'): the voicing assimilation rule (1c) applies in those cases as well. The generality of the application of such rules highlights the fact that they apply to abstract subsets of features, not to actual phonetic or acoustic objects.

The depth of abstract representations becomes clear when we consider how these rules are embedded within a fuller set of optional phonological processes involved in describing English sound systems in general (Chomsky & Halle, 1968). Consider the verbs 'mound' and 'mount'; in one of their allowed past tense pronunciations, they appear as 'mo¯"uDed' (long nasalized vowel preceding a tongue flap) and 'mo¯uDed' (short nasalized vowel preceding a tongue flap). These two words can end up being differentiated acoustically only in terms of the length of the first vowel, even though the underlying basis for the difference is in the voicing or absence of it in the final d/t of the stem (see Chomsky, 1964). The rules involved in arriving at these pronunciations include,

1d. nasalize a vowel before a nasal
1e. drop a nasal before a homorganic stop
1f. lengthen a vowel before a voiced consonant
1g. change t or d to a tongue flap, D, between vowels

Each of these rules has independent application in other parts of English pronunciation patterns, so they are distinct. If these rules are separated so

that they can apply to isolated cases, they must apply in order when combined. For example, since (1d) requires a nasal, (1e) cannot have applied; if (1f) is to apply differentially to 'mound', and not to 'mount', then (1e) must have applied; if the vowel length is to reflect the difference between 'mound' and 'mount', (1f) must apply before (1g). Thus, whether the intermediate stages are serially or logically ordered, the inputs, mound+past, mount+past, have a very abstract relation to their corresponding outputs:

mound+past		mount+past
mound D/T	(rule 1a)	mount D/T
mound eD/T	(rule 1b)	mount eD/T
mound ed	(rule 1c)	mount ed
m˜ounded	(rule 1d)	m˜ounted
m˜ouded	(rule 1e)	m˜outed
m˜ouded	(rule 1f)	(can't apply)
m˜ouDed	(rule 1g)	m˜ouDed

That the rules can be optional does not make them probabilistic within the model: they state what is allowed, not how often it happens (indeed, for many speakers, deleting nasals can occur more easily before unvoiced homorganic stops, than before voiced stops). Optionality in a rule is a way of expressing the fact that the structure can occur with and without a corresponding property.

The fact that linguistic rules apply to their appropriate domain 'without exception' does not mean that the appropriate domain is defined only in terms of phonological units. For example, in English there is a particular set of 'irregular' verbs which are not subject to the three past-tense formation rules described above. Whether a verb is irregular or not depends on its lexical representation: certain verbs are and others are not 'regular'. For example, one has to know which 'ring' is meant to differentiate the correct from incorrect past tenses below (see Pinker & Prince, 1988 for other cases).

2a. The Indians ringed (*rang) the settler's encampment.
2b. The Indians rang (*ringed) in the new year.

There are about 200 irregular verbs in modern English. They are the detritus of a more general rule-governed system in Old English (which have an interesting relation to the Indo-European e/o ablaut, see Bever & Langendoen, 1963). They fall into a few groups; those involving no change (which characteristically already end in t or d (beat, rid); those which add t (or d) (send); those lowering the vowel (drink, give); those involving a reversal of the vowel color between front and back (find, break; come); those which both lower and change vowel color (sting); those which involve combinations

of all three kinds of change (bring, feel, tell). The point for our purposes is that almost all of the 'irregular' verbs draw on a small set of phonological processes. Only a few involve completely suppletive forms (e.g., go/went).

This brief analysis of the past tense formation in terms of features and rules, reveals several properties of the structural system. First, the relevant grouping of features for the rules is vertical, with features grouped into phonemic segments. This property is not formally necessary. For example, rules could range over isolated features collected from a series of phones: apparently, it is a matter of linguistic fact that the phoneme is a natural domain of the phonological processes involved in the past tense formation (note that there may be suprasegmental processes as well, but these tend to range over different locations of values on the same feature dimension). Second, it is the segment at the end of the verb stem that determines the ultimate features of the regular past tense ending. This, too, is a fact about the English past system, not a logically necessary property.

4. The R&M model: A general picture of PDP models of learning

Rumelhart and McClelland (1986b; R&M) implemented a model which learns to associate past tense with the present tense of both the majority and minority verb types. The first step in setting up this model is to postulate a description of words in terms of individual feature units. Parallel distributed connectionist models are not naturally suited to represent serially ordered representations, since all components are to be represented simultaneously in one matrix. But phonemes, and their corresponding bundles of distinctive features, clearly are ordered. R&M solve this problem by invoking a form of phonemic representation suggested by Wickelgren (1969), which recasts ordered phonemes into 'Wickelphones', which can be ordered in a given word in only one way. Wickelphones appear to avoid the problem of representing serial order by differentiating each phoneme as a function of its immediate phonemic neighborhood. For example, 'bet' would be represented as composed of the following Wickelphones.

eT#, bEt, #Be

Each Wickelphone is a triple, consisting of the central phoneme, and a representation of the preceding and following phonemes as well. As reflected in the above representation, such entities do not have to be represented in memory as ordered: they can be combined in only one way into an actual sequence, if one follows the rule that the central phone must correspond to the prefix of the following unit and the postfix of the preceding unit. That rule leads to only one output representation for the above three Wickel-

phones, namely 'b...e...t'. Of course, the number of Wickelphones in a language is much larger than the number of phonemes—roughly the third power. But such a representational scheme seems to circumvent the need for a direct representation of order itself (at least, so long as the vocabulary is restricted so that a given Wickelphone never occurs more than once in a word—see Pinker & Prince, 1988).

Figure 2. *Categorization of phonemes on four simple dimensions*

		Place					
		Front		Middle		Back	
		V/L	U/S	V/L	U/S	V/L	U/S
Interrupted	*Stop*	b	p	d	t	g	k
	Nasal	m	–	n	–	N	–
Cont. Consonant	*Fric.*	v/D	f/T	z	s	Z/j	S/C
	Liq/SV	w/l	–	r	–	y	h
Vowel	*High*	E	i	O	ʌ	U	u
	Low	A	e	I	a/α	W	*/o

Key: N = ng in *sing*; D = th in *the*; T = th in *with*; Z = z in *azure*; S = sh in *ship*; C = ch in *chip*; E = ee in *beet*; i = i in *bit*; O = oa in *boat*; ʌ = u in *but* or schwa; U = oo in *boot*; u = oo in *book*; A = ai in *bait*; e = e in *bet*; I = i_e in *bite*; a = a in *bat*; α = a in *father*; W = ow in *cow*; * = aw in *saw*; o = o in *hot*.

R&M assign a set of phonemic distinctive features to each phone within a Wickelphone. There are 4 feature dimensions, two with two values and two with three, yielding 10 individual feature values (see Figure 2). This allows them to represent Wickelphones in feature matrices: for example the /E/ in 'bet' would be represented as shown below.

Dimension 1	Interrupted	Vowel	Interrupted
Dimension 2	Stop	Low	Stop
Dimension 3	Voiced	Short	Unvoiced
Dimension 4	Front	Front	Middle
	f1	f2	f3

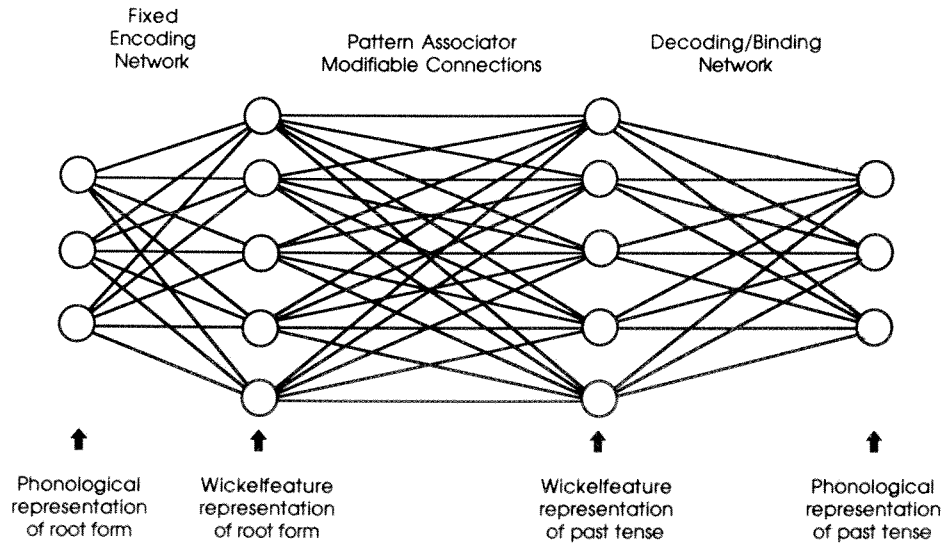
The verb learning model represents each Wickelphone in a set of 'Wickelfeatures'. These consist of a triple of features, [f1, f2, f3], the first taken from the prefix phone, the second from the central phone and the third from the post-fix phone. Accordingly, some of the Wickelfeatures for the 'bet' would be the following:

f1	f2	f3
[end,	interrupted,	vowel]
[end,	interrupted,	low]
[stop,	low,	stop]
[voiced,	low,	unvoiced]

There are about 1000 potential Wickelfeatures (10 prefix values × 10 central phone values × 10 post-fix values)—(the potential number is somewhat higher because of the treatment of word boundaries; we describe below how the actual number is limited).

There are two layers at which the Wickelfeature representations of the words occur, the input and the output. Each layer consists of a separate node for each of the Wickelfeatures: all of the nodes at each layer are connected to all of the nodes at the other, as depicted in Figure 3. The machine is taught in the following way: the input is provided in the form of a conventional phonemic notation, and transformed into a corresponding set of Wickelfeatures. The input layer of Wickelfeatures is activated by the input. Each input node is connected to each output node with a specific weight. On each trial, the weights of these connections determine the influence of the activation of the input node on the activation of the output node it is connected to. If the weight between an input node and an output node is 0, then the state of the

Figure 3. *(Reproduced from Rumelhart and McClelland, 1986b, p. 222, with permission of the publisher, Bradford Books/MIT Press.)*



input node does not affect the state of the output node. If the weight is positive, and if the input node is activated by the present-tense input, the output node is also activated by the connection between them; if the weighting is negative then the output node would be inhibited by that connection. Each output node also has a threshold, which the summed activation input from all input nodes must exceed for the output node to become active (we note below that the threshold is itself interpreted as a probabilistic function).

On each training trial, the machine is given the correct output Wickelfeature set as well as the input set. This makes it possible to assess the extent to which each output Wickelfeature node which should be activated, is, and conversely. The machine then uses a variant on the standard perceptron learning rule (Rosenblatt, 1962), which changes the weights between the active input nodes and the output nodes which were incorrect on the trial: lower the weight and raise the threshold for all nodes that were incorrectly activated; do the opposite for nodes that were incorrectly inactivated.

The machine was given a set of 200 training sessions, with a number of verbs in each session. At the end of this training, the system could take new verbs, that it had not processed before, and correctly associate their past tense in most cases. Hence, the model appears to learn, given a finite input, how to generalize to new cases. Furthermore, the model appears to go through several stages of acquisition which correspond to the stages of learning the past tense of verbs which children go through as well (Brown, 1973; Bybee & Slobin, 1982). During an early phase, the model (and children) produce the correct past tense for a small number of verbs, especially a number of the minority forms (went, ran, etc.). Then, the model (and children) 'overgeneralize' the attachment of the majority past form, 'ed' and its variants, so that they then make errors on forms on which they had been correct before (goed, wented, runned, etc.). Finally, the model (and children) produce the correct minority and majority forms. It would seem that the model has learned the rule-governed behaviors involved in forming the past tense of novel verbs. Yet, as R&M point out, the model does not 'contain' rules, only matrices of associative strengths between nodes. They argue that the success of the system in learning the rule-governed properties, and in simulating the pattern of acquisition, shows that rules may not be a necessary component of the description of acquisition behavior.

Whenever a model is touted to be as successful as theirs, it is hard to avoid the temptation to take it as a replacement for a rule-governed description and explanation of the phenomena. That is, we can go beyond R&M and take this model as a potential demonstration that the appearance that behavior is rule-governed is an illusion, and that its real nature is explained by nodes and the associative strengths between networks of nodes. A number

of linguistic commentators have drawn this conclusion about connectionist models, in particular because of the apparently successful performance of R&M's past-tense learning model (Langackre, 1987; Sampson, 1987). Below, we examine the model in this light: we find no evidence that the model replaces linguistic rules; rather, it has internal architecture which is arranged to be particularly sensitive to aspects of data that conform to the rules. Thus, the model may (or may not) be a successful algorithmic description of learning; but, insofar as it works, it actually confirms the existence of rules as the basis for natural language.

5. The model's TRICS (The Representations It Crucially Supposes)

We now turn to some special characteristics of the model which contribute to making it work. In each case, we note that apparently arbitrary decisions, often presented as 'simplifications' of the model, actually could be explained as accommodations to the rule-governed properties of the system and the resulting behavior. We should note that, by and large, it did not require much detective work to isolate the arbitrary properties—in most cases, R&M are explicit about them. There are two kinds of TRICS, those that reconstitute crucial aspects of the linguistic system, and those which create the overgeneralization pattern exhibited by children.

5.1.

A set of properties of the model serve to re-focus on the phoneme-by-phoneme clustering of features, and to emphasize the clarity of the information at the end of the verb. That is, the arbitrary decisions about the details of the model are transparently interpretable as making most reliable for associative learning, the information relevant to the rule-governed description of the formation of the past tense. Our forensic method is the following: we examine each arbitrary decision about the model with the rule system in mind, and ask about each decision: would this facilitate or inhibit the behavioral emergence of data which looks like that governed by the past tense rules? Without exception, we find that the decision would facilitate the emergence of such behavior.

5.1.1.

The first simplification of the model involves reducing the number of within-word Wickelfeatures from about 1000 to 260. One way to do this, would be to delete randomly some Wickelfeatures, and rely on the overall

redundancy of the system to carry the behavioral regularities. Another option is to use a principled basis for dropping certain Wickelfeatures. For example, one could drop all Wickelfeatures whose component subfeatures are on different dimensions—this move alone would reduce the number of features considerably. Such reduction would occur if Wickelfeatures were required to have at least two sub-features on the same dimension. R&M do something like this, but in an eccentric way: they require that all Wickelfeatures have the same dimension for f1 and f3: f2 can range fully across all feature dimensions and values. Accordingly, the potential Wickelfeatures for the vowel /E/ in 'bet' on the left below are possible, those on the right are not.

[interrupted, vowel, interrupted]	[interrupted, vowel, stop]
[voiced, mid, unvoiced]	[stop, vowel, unvoiced]
[front, short, middle]	[front, short, unvoiced]

This apparently arbitrary way of cutting down on the number of Wickelfeatures has felicitous consequences for the relative amount of rule-based information contained within each sub-feature. It has the basic effect of reducing the information contained within f1 and f3, since they are heavily predictable—that is the actually employed Wickelphones are "centrally informative." This heightens the relative importance of information in f2, since it can be more varied. This move is entirely arbitrary from the standpoint of the model; but it is an entirely sensible move if the goal were to accommodate to a structural rule account of the phenomena: the rules imply that the relevant information in a phoneme is in f2. The use of centrally informative Wickelphones automatically emphasizes f2.

5.1.2.

Word boundaries present a special descriptive problem. By virtue of being boundaries, most feature values are irrelevant. In standard phonology, word boundary is often treated as an environmental feature, without phonetic content of its own. The difficulty for a Wickelfeature notation is that word-boundary phones have an f1 or f3 which designates the word boundary. This requires that some sort of Wickelfeature status be given to f1, 2, 3 triples which have either f1 or f3 as a word boundary. A theoretically neutral solution to this would be to treat word-boundary as a bi-polar feature, and allow it to combine with other features into Wickelfeatures. R&M do something like this, but again in an eccentric way. They set up a completely separate set of 200 Wickelfeatures just for phones at word boundaries. Consider word-final phones. R&M allow the available Wickelfeatures to be the cross product of all possible values of f1 and f2, so long as f3 is the boundary. For example, all the features below are among the possible features for the /t/ in /bEt/.

[vowel, interrupted, end] [low, unvoiced, end]
[front, stop, end] [short, middle, end]

We can see that this gives a privileged informational status to phones at the word boundary, compared with any other ordinally defined position within the word: the phones at the boundary are the only ones with their own unique set of Wickelfeatures. This makes the information at the boundary uniquely recognizable. That is, the Wickelfeature representation of the words exhibits the property of 'boundary sharpening'.

This arbitrary move accommodates another property we noted in the rule-governed account of the past tense. The phone at the word boundary determines the shape of the regular past. Here, too, we see that an apparently arbitrary decision was just the right one to make in order to make sure that the system would accommodate to the rule-governed regularities.

5.1.3.

R&M allow certain input Wickelfeatures to be activated even when only some of the constituent sub-features are present. One theoretically arbitrary way to do this would be to allow some Wickelphone nodes to be activated sometimes when any 1 out of 3 subfeatures does not correspond to the input. R&M do something like this, but yet again, in an eccentric way—the model allows a Wickelphone node to be activated if either f1 or f3 is incorrect, but not if f2 is incorrect. That is, the fidelity of the relationship between an input Wickelphone and the nodes which are actually activated, is subject to 'peripheral blurring'. This effect is not small: the blurring was set to occur with a probability of .9 (this value was determined by R&M after some trial and error with other values). That is, a given Wickelnode is activated 90 percent of the time when either the input does not correspond to its f1 or f3. But it can always count on f2. This dramatic feature of the model is unmotivated within the connectionist framework. But it has the same felicitous result from the standpoint of the structure of the phenomenon as discussed in 5.1.1. It heightens (in this case, drastically) the relative reliability of the information in f2, and tends to destroy the reliability of information in f1 and f3. This further reflects the fact that the structurally relevant information is in f2.

Blurring, however, has to be kept under careful control. Clearly, if blurring occurred on 100 percent of the features with incorrect f1 or f3, the crucial information will be lost as to how to sequentially order the phones. For this reason, R&M had to institute an arbitrary choice of which Wickelfeatures could be incorrectly activated, so that for each input feature there are some reliable cues to Wickelphone order. This necessitates blurring the Wickelfeatures with less than 100 percent of the false options.

5.1.4.

The use of phonological features is one of the biggest TRICS of all. Features are introduced as an encoding device which reduces the number of internodal connections—the number of connections between two layers of 30,000 Wickelphones is roughly one billion. Some kind of encoding was necessary to reduce this number. Computationally, any kind of binary encoding could solve the problem: R&M chose phonological features as the encoding device because that is the only basis for the model to arrive at appropriate generalizations. Furthermore, the four distinctive features do not all correspond to physically definable dimensions. In order to simplify the number of encoded features, R&M create some feature hybrids. For example b,v,EE are grouped together along one 'dimension' in opposition to m,l,EY. While such a grouping is not motivated by either phonetics or linguistic theory, it is neither noxious nor helpful to the model, so far as we can tell. However, the other major arbitrary feature grouping lumps together long vowels and voiced consonants in opposition to short vowels and unvoiced consonants. All verbs which end in vowels, end in long vowels: accordingly, that particular grouping of features facilitates selective learning of the verbs that end in a segment which must be followed by /d/, namely the verbs ending in vowels and in voiced consonants.

5.1.5.

It is interesting to ponder how successful the devices of central informativeness, peripheral blurring and boundary sharpening are in reconstituting traditional segmental phonemic information. Phonemic representations of words offer a basis for representing similarity between them. One of the properties of Wickelphones is that they do not represent shared properties of words. For example, 'slowed' and 'sold' are no more similar in Wickelphones than 'sold' and 'fig'. This is obviously an inherent characteristic of Wickelphonology (Pinker & Prince, 1988; Savin & Bever, 1970) which would extend to Wickelfeature representations if there were no TRICS. But there are, and they turn out to go a long way to restoring the similarity metric represented directly in normal phonemic notations. One way to quantify this is to examine how many shared input Wickelfeatures there are for words which do and do not share phonemes. This is difficult to do in detail, because R&M do not give complete accounts of which features were chosen for blurring. Two arbitrarily chosen words without shared phonemes will also share few Wickelfeatures—our rough calculation for a pair of arbitrarily chosen 4 letter words is about 10 percent. Words which share initial or final phonemes, will have a noticeable number of shared features because of boundary sharpening. Blurring plays an especially important role in reconstituting similarity among

words with shared internal phonemes, such as 'slowed' and 'sold'. Roughly, our calculations show that two such words go from about 20 percent shared features without blurring to around 65 percent with it. In terms of correlating the chances of a node being activated in each word, this represents a rise from about .1 to about .5. The corresponding proportions for phonemically distinct words are 10 and 30 percent—the correlation in that case, stays at 0, even with blurring. The technical reason for this proportional difference is that the blurring has a radiating effect on just the right Wickelfeatures to create an overlap when there are common phonemes. Thus, the model does not correctly reconstitute the phonemic representation, but it does replicate some of its effects.

5.2.

There are two major behavioral properties of the model. First, it seems to learn (i.e., it changes its behavior); second, it goes through a period of overgeneralizing the regular rule. The fact that the model learns at all has the same formal basis as the fact that the model can represent the present → past mapping. Any mapping a perceptron can carry out, it can 'learn' to carry out, using the kind of learning rule described above (Rosenblatt, 1962). Basically, the model is composed of 460 perceptrons which converge on the appropriate mapping representation. In fact, in simple perceptrons, the convergence can be quite efficient. On each trial, the learning rule adjusts a threshold discrimination function such that only the correct output units are activated. What is striking about R&M's complex perceptron is that it takes so long to learn the mapping. The function for regular verbs is extremely simple, and might well be reasonably arrived at with just one training cycle. Thereafter, a few cycles would suffice to straighten out the errors on the irregular forms, especially because, as we pointed out above, most irregular verbs follow a restricted set of rules. In this case, there would be little intermediate performance characteristic of learning, and no period during which the regular endings overgeneralized to previously correct irregular verbs.

The reason that the model in fact does exhibit considerable intermediate performance and overgeneralization is due to another one of the TRICS, which imposes a probabilistic function on the output. The probability of an output unit being active is a sigmoid function of its net activation (weighted input minus threshold): this function represents the fact that even when the input is correctly assigned, there are output errors. Such a move qualitatively improves the generalizations made by the system once it has been trained. This is because, in general, as the number of error-correcting trials increases, the difference between activations resulting from inputs for which the unit is

supposed to have positive output and those for which it is supposed to have negative output, grows more rapidly than the difference among the inputs of each type. This enhances the clarity of the generalization and also makes the learning proceed more slowly. As R&M put it, its use here is motivated by the fact that it "causes the system to learn more slowly so the effect of regular verbs on the irregulars continues over a much longer period of time." (R&M, p. 224).

The period of overgeneralization of the regular past at the 11th cycle of trials also depends on a real trick, not a technically defined one. For the first 10 cycles, the machine is presented with only 10 verbs, 8 irregular and 2 regular ones. On the 11th cycle it is presented with an additional 410 verbs of which about 80 percent are regular. Thus, even on the 11th cycle alone, the model is given more instances of regular verbs than the training trials it has received on the entire preceding 10 cycles: it is no wonder that the regular past ending immediately swamps the previously acquired regularities. R&M defend this arbitrary move by suggesting that children also experience a sudden surge of regular past tense experience. We know of no acquisition data which show anything of the sort (see Pinker & Prince, 1988, who compile evidence to the contrary). Furthermore, if there were a sudden increase in the number of verbs a child knows at the time he learns the regular past tense rule, it would be ambiguous evidence between acquiring the rule, and acquiring a lot of verbs. The rule allows the child to memorize half as many lexical-items for each verb, and learn twice as many verbs from then on. Therefore, even if it were true that children show a sudden increase in the number of verbs they know at the same time that they start overgeneralizing, it would be very difficult to decide which was the cause and which the effect.

5.3. TRICS aren't for kids

It is clear that a number of arbitrary decisions made simply to get the model up and working, were made in ways that would facilitate learning the structural regularities inherent to the presented data. To us it seems fairly clear what went on: Wickelphones were the representation of choice because they seem to solve the problem of representing serial order (though they do so only for a restricted vocabulary, see Pinker & Prince, 1988; Savin & Bever, 1970). But Wickelphones also give equal weight to the preceding and following phone, while it is the central phone which is the subject of rule-governed regularities. Accordingly, a number of devices are built into the model to reduce the information and reliability of the preceding and following subphones in the Wickelphone. Further devices mark phones at word-boundary as uniquely important elements, as they are in rule-governed accounts of

some phonological changes which happen when morphemes are adjoined. Finally, the behavioral learning properties of the model were insured by making the model learn slowly, and flooding it with regular verbs at a particular point.

The most important claim for the R&M model is that it conforms to the behavioral regularities described in rule-governed accounts, but without any rules. We have not reported on the extent to which the model actually captures the behavioral regularities. Pinker & Prince (1988) demonstrate that, in fact, the model is not adequate, even to the basic facts: hence, the first claim for the model is not correct. We have shown further, that even if the model *were* empirically adequate, it would be because the model's architecture is designed to extract rule-based regularities in the input data. The impact of the rules for the past tense learning, is indirectly embedded in the form of representation and the TRICS: even Wickelfeatures involve a linguistic theory with acoustic segments and phonological features within them; the work of the TRICS is to render available the segmental phoneme, and emphasize boundary phonemes in terms of segmental features. That is, garbage in/garbage out: regularities in/regularities out. How crucial the TRICS really are is easy to find out: simply run the model without them, and see what it does. We expect that if the TRICS were replaced with theoretically neutral devices, the new model would not learn with even the current limited 'success', if at all; nor would it exhibit the same behaviors.

If a slightly improved set of TRICS does lead to successful performance, one could argue that this new model is a theory of the innate phonological devices available to children. On this interpretation, the child would come to the language learning situation with uncommitted connectionist networks supported by TRICS of the general kind built into the R&M model. The child operates on the feedback it receives from its attempts to produce phonologically correct sequences, and gradually builds up a network which exhibits rule-like properties, but without any rules, as R&M claim. It is difficult to consider the merits of such a proposal in the abstract: clearly, if the TRICS were sufficiently structured so that they were tantamount to an implementation of universal phonological constraints in rule-governed accounts, then such a theory would be the equivalent of one that is rule-based (see Fodor & Pylyshyn, 1988, for a discussion of connectionist models as potential implementation systems). The theory in R&M self-avowedly and clearly falls short of representing the actual rules. So, we must analyze the nature of the TRICS in the model at hand, to assess their compatibility with a plausible universal theory of phonology.

None of the TRICS fares well under this kind of scrutiny. Consider first the limitation on Wickelfeatures which makes them 'centrally informative':

this requires that f1 and f3 be marked for the same feature dimension (although the values on that dimension may be different). Certain phonological processes depend on information about the phone preceding and following the affected segment. For example, the rule which transforms /T or D/ to a voiced tongue flap, applies only when both the preceding and following segments are vowels. The 'central-informativeness' TRIC neatly accommodates a process like this, since it makes available a set of Wickelfeatures with f1 and f3 marked as 'vowel'. Unfortunately, the same TRIC makes it hard to learn processes in which f1 and f3 are marked for different dimensions. Since such processes are also quite common, the universal predictions this makes are incorrect.

The second set of representational TRICS has the net result of sharpening the reliability of information at word boundaries: this is well-suited to isolate the relevant information in the regular past-tense formation in English, and would seem like a natural way to represent the fact that morphological processes effect segments at the boundaries of morphemes when they are combined. Unfortunately, such processes do not seem to predominate over within-word processes, such as the formation of the tongue-flap between vowels, or the restrictions on nasal deletion. Furthermore, there are numerous languages which change segments within morphemes as they are combined, as in the many languages with vowel harmony. Thus, there is no empirical support for a system which unambiguously gives priority to phonological processes at morpheme boundaries.

R&M link together distinctive features which are maintained orthogonal to each other in most phonological theories. Hence, in R&M, long vowels and voiced consonants are linked together in opposition to short vowels and unvoiced consonants. This link is prima facie a TRIC, which facilitates learning the regular past tense morphology. But, taken as a claim of universal phonological theory, it is prima facie incorrect: it would propose to explain a non-fact, the relative frequency, or ease of learning, processes which apply simultaneously to long vowels and voiced consonants or to short vowels and unvoiced consonants.

Staging the input data, and imposing a sigmoid learning function are not, strictly speaking, components of phonological theory—both devices play a role in guaranteeing overgeneralization of the regular past. Such overgeneralizations are common in mastery of other morpho-phonological phenomena, for example, the present tense in English ('Harry do-es' (rhymes with 'news')) or plural ('fishes', 'childrens'). The carefully controlled sigmoid learning function and the staging of input data necessary to yield overgeneralization phenomena, have the properties of dei ex machina—with no independent evidence of any kind.

In brief, the net effect of the TRICS is to refocus the reliable information within the central phone of Wickelphone triples. This does reconstitute the segmental property of many phonological processes, obscured by the Wickelphonological representations. However, since those representations are not adequate in general, such reconstitution is of limited value. Most important, the specific TRICS involved in this reconstitution make wrong universal claims in some cases, and make obscure and unmotivated claims in the other cases. We conclude that even if the TRICS were fine-tuned to arrive at satisfactory empirical adequacy, they would still be arbitrarily chosen to facilitate the learning of English past tense rules, with no empirical support as the basis for phonological learning in general.

We noted above that there are theories of phonology in which more than one segment contributes to a constraint simultaneously, e.g., 'auto-segmental phonology' (Goldsmith, 1976). It might seem that such a variant of phonological theory would be consistent with Wickelphones—and the connectionist learning paradigms in general. Such claims would be incorrect. First, autosegmental phonology does not deny that many processes involve individual segments; rather, it asserts that there are simultaneous suprasegmental structures as well. Second, Wickelphones are no better suited than simple phones for the kinds of larger phonological unit to which multi-segmental constraints apply, very often the syllable. Finally, the constraints in autosegmental phonology are structural and just as rule-like as those in segmental phonology. It might also seem that the issue between traditional and autosegmental phonological theory concerns the reality of intermediate stages of derivation as resulting from the ordered application of rules like (1a–g)—another way to put this is in terms of whether rules are simple-but-ordered as in traditional phonology, or complex-but-unordered. There may be phonological theories which differ on these dimensions. But, however this issue is resolved, there will be no particular comfort for connectionist learning models of the type in R&M. The underlying object to which the rules apply will still be an abstract formula, and the output will still differentiate categorically between grammatical and ungrammatical sequences in the particular language.

5.4. Empirical evidence for rules

Up to now, we have relied on the reader's intuitive understanding of what a 'rule' is—a computation which maps one representation onto another. We have argued further that the R&M model achieves categorical rule-like behavior in the context of an analogue connectionist machine by way of special representational and processing devices. One might reply that the 'rules' we have been discussing actually compute the structure of linguistic 'compe-

tence', while the TRIC-ridden model is a 'performance' mechanism which is the real basis for the rule-like behavior. This line of reply would be consistent with the current distinction between three types of description: the computational, the algorithmic and the implementational (Marr, 1982). It is a line already taken in several defenses of connectionism (Rumelhart & McClelland, 1986a; Smolensky, in press).

The distinction between these different types of description might seem to allow for a synthesis of the connectionist and rule-based theories. On this view, rule-based theories describe the structure of language, while connectionist models explain how it 'actually' works. This would-be synthesis is not available, however, since grammatical rules are necessary for the explanation of behavior.

5.4.1. The diachronic maintenance of language systems

Consider first the operation of the processes we have used as examples: they characteristically fall into categories, often even at a physical level of description. For example, if a stop sound is 'unvoiced', it exhibits certain invariants which contrast it from its 'voiced' mode. The indicated processes occur in environments which are categorically described—e.g., /t,d/ becomes a tongue flap between two vowels (actually between two 'non-consonants' in distinctive feature terms), not between two sounds that are like vowels to a high degree. Variations in language behavior show similar discontinuities; for example, children invent phonological rules which involve rule-governed shifts rather than just groups of changed words. Similarly, dialects differ by entire rule processes, not isolated cases; finally, stable historical changes occur in precise but broad ranging shifts—the great vowel shift involved in the irregular past verbs included a complete rotation of vowel heights, not isolated changes. We are not suggesting that developmentally, synchronically and historically, there are no intermediate stages of performance; rather, we emphasize that the stable phenomena and periods are those caused by mental representations that are structural in nature.

It is possible to show that mental representations of a rule-based account of language are necessary to describe the properties of language change. The categorical nature of language change is explained in a rule-based account by the fact that rules themselves are categorical, not incremental: hence, linguistic change is resisted except at those times when it occurs in major shifts. Such facts are clearly consistent with rules, but it must be shown that they are consistent with models like that in R&M. A way to do this is to consider whether successive generations of such models could maintain an approximation of rule-governed behavior, without containing rules. The models we have considered achieve 90–95 percent correct output on their training set

and considerably less on generalization trials (Pinker & Prince, 1988, calculate 66 percent correct on generalizations). This level is, of course, far below a 5-year-old child's ability, but one might argue that improved models will do better. However, if these models are to be taken as anything like correct models of the child, they must exhibit stable as well as accurate behavior, in the face of imperfect input. We can operationalize this by asking a simple question (or performing the actual experiment): what will an untrained model learn, if it is given as input, the less-than-perfect output of a trained model?

This question is divisible into two parts: how fast will the 'child' model arrive at its asymptotic performance compared with the 'parent', and what will the asymptotic level be? It is likely that for a given number of trials before asymptote is reached, the child-model will perform worse than the parent-model. This follows from the fact that the data the child-model is given are less reliably related to the actual structure of the language, and therefore must require more trials to arrive at a stable output.

It is less clear what the final asymptotic level will be. If the parent-model errors were truly random in nature, then the final asymptotic level of performance should be the same in the child model. But, in fact, the parental errors are *not* random—they tend to occur on just those forms which are hard for the model to learn. R&M offer a case in point: after 80,000 trials, the model still makes a variety of strange errors on the past tense (e.g., 'squawked' for 'squat'; 'membled' for 'mail'; see Pinker & Prince, 1988, especially for an analysis of the 'blending' mechanism which produces cases like the second). It is intuitively clear, that some of these errors occur because of phonological coincidences, others because of the overwhelming frequency of the regular past ending. In both kinds of cases, the errors have a systematic basis, and are not random—indeed, they are by operational definition, just the cases which frequency-based computations in both models discriminate with difficulty: so, we can expect the child-model to perform even worse on these cases, once given seductively misleading analyses by the parent model. Eventually, with some number of generations (itself determined by the learning curve parameters, and other TRICS), the final descendant-model will stabilize at *always* getting the critical cases wrong.

There are many indeterminacies in these considerations, and the best way to see what happens will be to train successive generations of models. We think that this is an important empirical test for any model of learning. One must not only show that a particular model can approximate rule-like behavior, given perfect input and perfect feedback information, but that successive generations of exactly the same kind of model continue to re-generate the same rule-like regularities, given the imperfect input of their immediate ancestor-models. R&M considered in this light, predicts that in successive gen-

erations, a language system will degenerate quickly towards the dominant rule, overgeneralizing most of the exceptional cases. But, this does not occur in actual linguistic evolution. Rather, it is characteristic that every systematic process has some sub-systematic exceptions. As Sapir put it, 'grammars always leak'. One can speculate as to why this is so (Bever, 1986; Sadock, 1974). The fact that it is so poses particular problems for a model based on imperfect frequency approximations by successive generations.

We do not doubt that a set of diachronic TRICS can be tacked onto the model, which would tend to maintain the rule-like regularities in the face of systematically imperfect input information. We expect by induction on the properties of the current TRICS, that two things will be true about any new TRICS: (1) insofar as they have any systematic motivation, it will not be from the connectionist framework, but from the rule-based explanation; (2) insofar as they work, it will be because of their relation to the rule-based account.

5.4.2. Linguistic intuitions

It is also striking that children seem to have explicit differentiation of the way they talk from the way they should talk. Children are both aware that they overgeneralize, and that they should not do it. Bever (1975) reports a dialogue demonstrating that his child (age 3;6) had this dual pattern.

Tom: Where's mommy?
Frederick: Mommy goed to the store.
Tom: Mommy goed to the store?
Frederick: NO! (annoyed) Daddy, *I* say it that way, not you.
Tom: Mommy *wented* to the store?
Frederick: No!
Tom: Mommy went to the store.
Frederick: That's right, mommy wennn ... mommy goed to the store.

Slobin (1978) reported extended interviews with his child demonstrating a similar sensitivity: 'she rarely uses some of the [strong] verbs correctly in her own speech; yet she is clearly aware of the correct forms.' He reports the following dialogue at 4;7.

Dan: ... Did Barbara read you that whole story ...
Haida: Yeah ... and ... mama this morning after breakfast, read ('red') the whole book ... I don't know when she readed ('reeded') ...
Dan: You don't know when she what?
Haida: ... she readed the book ...
Dan: M-hm
Haida: That's the book she read. She read the whole, the whole book.

Dan: That's the book she readed, huh?
Haida: Yeah ... *read*! (annoyed)
Dan: Barbara readed you Babar?
Haida: Babar, yeah. You know cause you readed some of it too ... she readed all the rest.
Dan: She read the whole thing to you, huh?
Haida: Yeah, ... nu-uh, you read some.
Dan: Oh, that's right; yeah, I readed the beginning of it.
Haida: Readed? (annoyed surprise) Read!
Dan: Oh, yeah — read.
Haida: Will you stop that Papa?
Dan: Sure

What are we to make of Frederick's and Haida's competence? On the one hand, they clearly made overgeneralization errors; on the other hand, they clearly knew what they should and should not say. This would seem to be evidence that the overgeneralizations are strictly a performance function of the talking algorithm, quite distinct from their linguistic knowledge. The fact that the children know they are making a mistake emphasizes the distinction between the structures that they know and the sequences that they utter. (But see Kuczaj, 1978, who showed that children do not differentiate between experimentally presented correct and incorrect past tense forms. We think that he underestimates the children's competence because of methodological factors. For example, he assumes that children think that everything they say is grammatical, which the above reports show is not true. Finally, in all his studies, the child in general prefers the correct past forms for the irregular verbs.)

In brief, we see that even children are aware that they are following (or should follow) structural systems. Adults also exhibit knowledge of the contrast between what they say and what is grammatical. For example the sentence below is recognized as usable but ungrammatical, while the second is recognized as grammatical but unusable (see discussion in Section 7 below).

Either I or you are crazy.

Oysters oysters oysters split split split

Children and adults who know the contrast between their speech and the correct form have a representation of the structural system. Hence, it is of little interest to claim that a connectionist model can 'learn' the behavior without the structural system. Real people learn both; most interestingly, they sometimes learn the structure before they master its use.

5.4.3. Language behaviors

There is also considerable experimental evidence for the independent role of grammatical structures in the behavior of adults. We discuss this under the rubric of evidence for a 'psychogrammar' (Bever, 1975), an internalized representation of the language, that is not necessarily a model of such behaviors as speech perception or production, but a representation of the structure used in those and other language behaviors. Presumably, the psychogrammar is strongly equivalent to some correct linguistic grammar with a universally definable mental and physiological representation. We set up the concept for this discussion to avoid claiming "psychological" or "physiological reality" for any *particular* linguistic grammar or mode of implementation. Rather, we wish to outline some simple evidence that a psychogrammar exists: this demonstration is sufficient to invalidate the psychological relevance of those connectionist learning models which do not learn grammars.

The fundamental mental activity in using speech is to relate inchoate ideas with explicit utterances, as in perception and production. There is observational and experimental evidence that multiple levels of linguistic representation are computed during these processes (Bever, 1970; Dell, 1986; Fodor, Bever, & Garrett, 1974; Garrett, 1975; Tanenhaus, Carlson, & Seidenberg, 1985). The data suggest that the processes underlying these two behaviors are not simple inversions, so they may make different use of the grammatical structures, suggesting separate representations of the grammar. Hence, the psychogrammar may be distinct from such systems of speech behavior; in any case, it explains certain phenomena in its own right. In standard linguistic investigations, it allows the isolation of linguistic universals due to psychogrammatical constraints from those due to the other systems of speech behavior. We think that the achievements of this approach to linguistic research have been prodigious and justify the distinction in themselves. A further argument for the separate existence of a psychogrammar is the empirical evidence that it is an independent source of acceptability intuitions. The crucial data are sequences which are intuitively well-formed but unusable, and sequences which are usable but intuitively ill-formed, as discussed above. Such cases illustrate that behavioral usability and intuitive well-formedness do not overlap completely, suggesting that each is accounted for by (at least partially) independent mental representations.

5.4.4. Conclusion: The explanatory role of rules

The evidence we have reviewed in the previous three sections demonstrates that even if one were to differentiate structural rules from algorithmic rules, it remains the case that the structural rules are directly implicated in the explanation of linguistic phenomena. That is, the rules are not merely abstract

descriptions of the regularities underlying language behaviors, but are vital to their explanation because they characterize certain mental representations or processes. It is because of those mental entities that the rules compactly describe facts of language acquisition, variation, and history; they provide explanations of linguistic knowledge directly available to children and adults; they help explain those mental representations involved in the comprehension and production of behaviorally usable sentences; they are part of the explanation of historical facts about languages.

6. Learning to assign thematic roles

We now turn to a second model in which linguistic behavior is apparently learned—the assignment of thematic roles to nounphrases in different serial positions (McClelland & Kawamoto, 1986, M&K). The system which learns to assign thematic roles to nouns in specific sentences has similar properties to the model which learns past tenses. The input representational nodes are triples, consisting of a syntactic position, and two semantic features for a noun or verb—the output nodes represent a semantic feature for a noun, one for a verb, and a thematic noun-verb relation. There is a probabilistic blurring mechanism, which turns on feature/role nodes only 85 percent of the time when they should be on, and 15 percent of the time when they should not.

The semantic TRIC

Ideally (that is, in one's idealization of how this model must work to be a significant psychological theory of learning to attach thematic roles to words), one would start with an independently defined set of semantic features (human, animate, ...) taken from some semantic theory (e.g., a theory intended to account for naming behavior, or within a semantic theory, to account for synonymy and entailment). Then, the role of statistical blurring might be interpreted as allowing for some interaction between these formal features and the continuous variability which can occur when fitting nouns into thematic roles. But, for all the statistical blurring, it remains the case that the 'semantic role features', do not flow from some independent theory: rather, they are just descriptors of roles themselves. Here, the semantic features are chosen for each noun to reflect the probability that it is an agent/object/instrument/modifier. The corresponding features for verbs are chosen to capture the likelihood that the verb is an action, involves modifiers and so on (see Figure 4). Hence, any 'learning' that occurs is trivial. The 'learning' does not involve isolating independently defined semantic features which are

Figure 4. *Feature dimensions and values*

	Nouns
HUMAN	human nonhuman
SOFTNESS	soft hard
GENDER	male female neuter
VOLUME	small medium large
FORM	compact 1-D 2-D 3-D
POINTINESS	pointed rounded
BREAKABILITY	fragile unbreakable
OBJ-TYPE	food toy tool utensil furniture animate nat-inan
	Verbs
DOER	yes no
CAUSE	yes no-cause no-change
TOUCH	agent inst both none AisP
NAT_CHNG	pieces shreds chemical none unused
AGT_MVMT	trans part none NA
PT_MVMT	trans part none NA
INTENSITY	low high

Note: nat-inan = natural inanimate, AisP = Agent is Patient, NA = not applicable.

relevant to roles, but rather an accumulation of activation strengths from having the role-features available, and being given correct instances of words (feature matrices) placed in particular role positions.

One of the achievements of this model according to M&K is that it 'over-generalizes' thematic role assignments. For example, 'doll' is not marked as 'animate' and therefore is ineligible to be an agent. However, 'doll' is nonetheless assigned significant strength as agent in such sentences as 'the doll moved'. This result seems to be due to the fact that everything is assigned the gender neuter except animate objects and the word 'doll' which is assigned 'female'. Thus, 'neuter' becomes a perfect predictor of inanimacy, except for 'doll'. It is not surprising that 'doll' is treated as though it were animate.

7. The power of units unseen

It might seem that the models we have discussed are dependent on TRICS because of some inherent limitation on their computational power. We now consider connectionist learning models with more computational power, and examine some specific instances for TRICS.

Connectionist learning machines are composed of perceptrons. One of the staple theorems about perceptrons is that they cannot perform certain Boolean functions if they have only an input and an output set of nodes (Minsky & Papert, 1969). Exclusive disjunctive conditions are among those functions that cannot be represented in a two-layer perceptron system. Yet, even simple phonetic phenomena involved in the past tense involve disjunctive descriptions, if one were limited to two-level descriptions. For example, the variants of 'mounded' discussed above involve disjunction of the presence of n, lengthened vowel and the tongue flap. That is, the tongue-flap pronunciation of 't' or 'd' can occur only if the 'n' has been deleted, and the previous vowel has been lengthened. Furthermore, the distinction between /t/ and /d/ in 'mounted' and 'mounded', in some pronunciations has been displaced to the length of the preceding vowel. The solution for modelling such disjunctive phenomena within the connectionist framework is the invocation of units that are neither input nor output nodes, but which comprise an intermediate set of units which are 'hidden' (Hinton & Sejnowski, 1983, 1986; Smolensky, 1986). (A formal problem in the use of hidden units is formulating how the perceptron learning rule should apply to a system with them: there are two (or more) layers of connections to be trained on each trial, but only the output layer is directly corrected. Somehow, incorrect weights and thresholds must be corrected at both the output and hidden levels. A current technique, 'back-propagation' is the instance of such a learning rule used in the examples we discuss below—it apportions 'blame' for incorrect activations to hidden units which are involved, according to a function too complex for presentation here (see Rumelhart, Hinton & Williams, 1986).)

Recent work has shown that a model with hidden units can be trained to regenerate a known acoustic sample of speech, with the result that novel speech samples can also be regenerated without further training (Elman & Zipser, 1986). The training technique does not involve explicit segmentation of the signal, nor is there a mapping onto a separate response. The model takes a speech sample in an input acoustic feature representation: the input is mapped onto a set of input nodes, in a manner similar to that of McClelland and Elman. Each input node is connected to a set of hidden nodes, which in turn are connected to a layer of output nodes corresponding to the input nodes. On each trial, the model adjusts weights between the layers of nodes

following the learning rule (the 'back-propagation' variant of it) to improve the match between the input and the output. This model uses an 'auto-associative' technique, in which weights are adjusted to yield an output that is the closest fit to the original input. After many trials (up to a million), the model is impressively successful, in regenerating new speech samples from the same speaker. This is an exciting achievement, since it opens up the possibility that an analysis of the speech into a compact internal representation is possible, simply by exposure to a sample. There are several things which remain to be shown. For example, the internal analysis which the model arrives at may or may not correspond to a linguistically relevant analysis; if it does correspond to such units, it is not clear how they can be integrated with higher-order relations between them.

7.1. A hidden unit model of syntax acquisition

Hanson and Kegl (1987, H&K) use the 'auto-association' method with hidden units to re-generate sequences of syntactic categories which correspond to actual sentences. After a period of learning, the model can take in a sequence of lexical categories (e.g., something like 'determiner, noun, verb, determiner, noun, adverb'), and regenerate that sequence. What is interesting is that it can regenerate sequences which correspond to actual English sentences, but it does not regenerate sequences which do not correspond to English sequences—in this way, the model approximates the ability to render grammaticality distinctions. Hanson and Kegl disavow their model as appropriate for the language learning child, but they make an extraordinarily strong claim for what it shows about linguistic structures which govern the ungrammaticality of certain sentences:

> If [our model] does not recognize such sentences ... after nothing more than exposure to data, this would lead us to suspect that rather than being an innate property of the learner, these constraints and conditions follow directly from regularities in the data ... . Both [our model] and the child are only exposed to sentences from natural language, they both must induce general rules and larger constituents from just the regularities to which they are exposed ... .

That is, H&K take the success of their model to be an existence proof that some linguistic universals can be learned without any internal structure. This makes it imperative to examine the architecture of their model—as we shall see, it incorporates certain linguistically defined representations in crucial ways which invalidate their empiricist conclusion.

Here is how one of the models works. The model is trained on a set of 1000 actual sentences, ranging from a few to 15 words in length. Each lexical

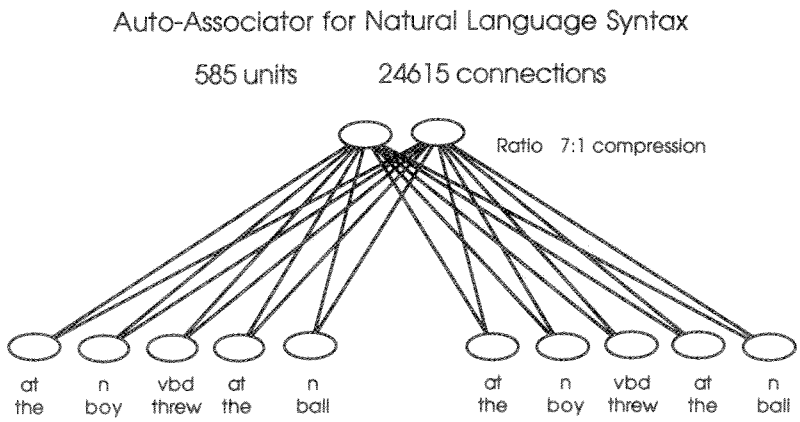
item in every input sentence is manually assigned to a syntactic category, each coded into a 9-bit sequence (input texts were taken from a corpus with grammatical categories already assigned: Francis & Kucera, 1979). These sequences are mapped onto 270 input nodes (135 for 15 word positions, each with nine bits; another distinct set for word boundary codes). The categorized sequences are then treated as input to a set of 45 hidden nodes—each input category node is connected to each hidden node, and each hidden node is connected to a corresponding set of 270 output nodes (see Figure 5). During training, the model is given input sequences of categories—the model matches the input against the self-generated output on each trial. The usual learning rule applies to adjust weights on each trial (using a variation of back-propagation), with the usual built-in variability in learning on each trial. After 180,000 trials with the training set, the model asymptoted at about 90 percent correct on both the training set and on new sentence-based category sequences which had not been presented before.

H&K highlight four qualitative results in the trained model's responses to new cases. First, the model supplements incomplete information in input sequences so that the regenerated sequences conform to possible sentence types. For example, given the input in (3a) the model's response fills in the missing word with a verb, as in (3b). (We are quoting directly from their examples. Roughly, the lexical categories correspond to distinctions used in Francis & Kucera, 1979. 'P-verb' refers to 'verb in past-tense form'.)

Second, the model corrects

3a. article, noun, ⟨BLANK⟩, article, noun
3b. article, noun, p-verb, article, noun

Figure 5.



incorrect input syntactic information; for example given (4a) as input, it responds with (4b).

4a. article, noun, p-verb, adverb, article, noun, p-verb
4b. article, noun, p-verb, preposition, article, noun, p-verb

The interest of this case is based on the claim that (4a) does not correspond to a possible sequence (e.g., they say (4a) corresponds to *'the horse raced quickly the barn fell', while (4b) corresponds to 'the horse raced past the barn fell'. Note that (4b) is a tricky sentence: it corresponds to 'the horse *that was* raced past the barn, fell').

Third, the model regenerates one center-embedding (corresponding to 'the rat the cat chased died'): this regeneration occurs "despite the lack of even a single occurrence of a center-embedded sentence within the [1000 sentence] corpus." Furthermore, the model rejects sequences corresponding to two center-embeddings ('the rat the cat the dog bit chased died'). Given a syntactic sequence corresponding to a double embedding (5a), the model responds with (5b). H&K say that this shows that their model "can differentially generalize to sentences that can appear in natural language (center-embeddings) but cannot recognize sentences which violate natural language constraints (multiple center-embeddings)."

5a. article, noun, article, noun, article, noun, p-verb, p-verb, p-verb
5b. article, noun, article, noun, article, noun, p-verb, noun, verb

Finally, the model refuses to regenerate sequences in which an adverb interrupts a verb and a following 'article noun' which would have to be the verb's direct object; for example given (6a) as input (corresponding to *'John gave quickly the book'), the model responds with (6b) (corresponding to 'John quickly was the winner'). H&K say that this shows that the model has acquired one of the universal case-marking constraints as expressed in English; "a direct object must be adjacent to a verb in order to receive case from it and thereby be allowed (licensed) to occur in object position ([from] ... a Government-Binding approach (Chomsky, 1981))."

6a. noun, verb, adverb, article, noun
6b. noun, adverb, was, article, noun

It is enterprising of H&K to put the model to such qualitative tests, over and above the 90 percent level of correct regenerations. As in formal linguistic research, it is the descriptive claims made by a model which justify its acceptance as much as its empirical adequacy to generate correct sentences. Unfortunately, the tests which H&K cite do not provide crucial support for their model, for a variety of reasons. First, many kinds of pattern recognition

models would isolate the fact that every sentence contains a verb—in their case, one of a set of pre-categorized input items, such as 'p-verb, was, do, etc ...'. It is not trivial that the model fills in a blank word as a sentence's only verb, but it is not unique either. Similarly, one would not be surprised if the model filled in a 'noun' for a blank between 'article' and 'verb', or an article or preposition for a blank between a 'verb' and a 'noun'.

The logic of the interpretation of the next two cases is contradictory. H&K cite with approval the fact that the model rejects a sequence corresponding to 'the horse raced breathtakingly the crowd roared' and corrects it to one corresponding to 'the horse raced past the barn fell'. Indeed, it is relevant that the model does not accept the input sequence, which is an obverse part of its 90 percent success in regenerating correct input sequences. In this case, the model responds by changing one lexical category so that the output corresponds to a possible English sentence—the question is, does it change the input to a correct or behaviorally salient sentence. Consider a sequence other than (4b) which would result from changing one category in (4a).

4c. art, noun, p-verb, conj, art, noun, p-verb ('the horse raced and the crowd roared')

Yet the model apparently chose (4b). This sentence type has long been understood as an example of a sentence which is *difficult* for speakers to understand (see Bever, 1970, for discussion). Accordingly, we take the fact that the model regenerates this behaviorally difficult sequence to be an empirical failure of the model, given other options like (4c) which correspond to much easier structures. Finally, the output in (4b) does not correspond to a well-formed sentence anyway. In the Francis et al. categorization schema H&K used to categorize their input, 'past verb' is differentiated from 'past participle', so the correct category sequence corresponding to 'the horse raced past the barn fell', would be as in (4d):

4d. article, noun, verb (past participle), article, noun, p-verb

The treatment of multiple center-embedding is conversely puzzling. Having taken the alleged success of the model at regenerating a behaviorally difficult sequence like (5a), H&K report approvingly that the model rejects doubly embedded sentences. Although they note that others have argued that such constructions are complex due to behavioral reasons, they appear to believe that in rejecting them, the model is simulating 'natural language constraints'. Others have also argued that the difficulty of center-embedded constructions shows that an adequate model of language structure should not represent them (e.g., McClelland & Kawamoto, 1986; Reich, 1969). For example, McClelland and Kawamoto write:

> The unparsability of [doubly-embedded] sentences has usually been explained by an appeal to adjunct assumptions about performance limitations (e.g., working-memory limitations), but it may be, instead, that they are unparsable because the parser, by the general nature of its design is simply incapable of processing such sentences.

There is a fundamental error here. Multiple center-embedding constructions are not ungrammatical, as shown by cases like (6c) (the unacceptability of (6d) shows that the acceptability of (6c) is not due to semantic constraints alone).

6c. The reporter everyone I have met trusts, is predicting another Irangate.
6d. The reporter the editor the cat scratched fired died.

In fact, the difficulty of center-embedding constructions is a function of the differentiability of the nounphrases and verbphrases: (6c) is acceptable because each of the nounphrases is of a different type, as is each of the verbphrases: conversely, (6d) is unacceptable because the three nounphrases are syntactically identical, as are the three verbphrases. This effect is predicted by any comprehension mechanism which labels nounphrases syntactically as they come in, stores them, and then assigns them to verbphrase argument positions as they appear: the more distinctly labelled each phrase is, the less likely it is to be confused in immediate memory (see Bever, 1970; Miller, 1962; for more detailed discussions). Thus, there are examples of acceptable center-embedding sentences, and a simple performance theory which predicts the difference between acceptable and unacceptable cases. Hence, H&K are simply in error to claim that it is an achievement of their model to reject center-embeddings, at least if the achievement is to be taken as reflecting universal structural constraints on possible sentences.

The other two qualitative facts suggest to H&K that their model has developed a representation of constituency. The model regenerates single embeddings without exposure to them, and rejects sequences which interrupt a verb-article-noun sequence with an adverb after the verb. Both behaviors indicate to H&K that the model has acquired a representation corresponding to a nounphrase constituent. They buttress this claim with an informal report of a statistical clustering analysis on the response patterns of the hidden units to input nounphrases and verbphrases: the clusters showed that some units responded strongly to nounphrases, while others responded strongly to verbphrases.

This seems at first to be an interesting result. But careful analysis of the internal structure of the model and the final weighting patterns are required to determine how it works. The grouping of those sequences which are nounphrases (sequences containing an article and/or some kind of noun) from

those which are verbphrases (containing one of the list of verb types) might occur for many reasons. Indeed, since verbphrases often contain nounphrases, it is puzzling that verbphrases did not excite nounphrase patterns at the same time: the very notion of constituency requires that they should. As it stands, the model appears to have encoded something like the following properties of 2–15 word sentences: they have an initial set of phrase types, and then another set of phrase types beginning with some kind of verb. This is an achievement, but not one that exceeds many item and arrangement schemata.

Hanson and Kegl's brief presentation (imposed on them by the publication format) does not allow us to examine the model completely for TRICS. We can note, however, several factors which may be important. The most serious issue involves the informativeness of the input categories, which are assigned by hand. The input is richly differentiated for the model into 467 syntactic categories. There are many distinctions made among the function words and morphemes. For example, many forms of 'be' are differentiated, personal possessive pronouns are differentiated from other pronouns, subject pronouns are differentiated from object pronouns, comparative adjectives are differentiated from absolutes, and so on.

Given a rich enough analysis of this kind, every English sentence falls into one kind of pattern or another. Indeed, language parsers can operate surprisingly successfully with differential sensitivity to 20 classes of function words and no differentiation of content words at all. The reason is straightforward: function words are the skeleton of syntax. They tend to begin English phrases, not end them: many function words offer unique information: e.g., 'the', 'a', 'my' always signal a noun somewhere to the right, 'towards' always signals a nounphrase or the end of a clause, and so on. In fact, it would be interesting to see if H&K's parser performs any worse if it is trained only on function word categorized input. As it stands, recognition of a few basic redundancies might well account for the model's 85 percent hit rate.

H&K state that their model 'begins with no assumptions about syntactic structure nor any special expectations about properties of syntactic categories other than the fact they exist.' This is a bit misleading. First, syntactic categories are not independently distinguished from the syntax in which they occur. Many, if not all, syntactic categories in a language are like phonological distinctive features, in that they are motivated in part by the rule-based function they serve universally and in a particular language. For example, in English, the words, 'in, on, under ...' are all prepositions just because they act the same way in relation to linguistic constraints—i.e., they precede nounphrases, can be used as verb-particles, can coalesce with verbs in the passive form, and so on. Giving the model the information that these privileges of

occurrence coincide with a category is providing crucial information about what words can be expected to pattern together—something which a real learner would have to discover. Thus, providing the correct categories provides information which itself reflects syntactic structure.

H&K actually give the model even more information: they differentiate members of the same category when they are used in syntactically different ways. For example, the prepositions are pre-categorized differently according to whether they are used in prepositional phrases or as verb particles. In the categorized samples they provide, the prepositions below on the left are classified with the symbols on the right:

to working	'in'
on the	'in'
to the	'in'
over the	'in'
pushed *aside*	'rb'
pass it *up*	'rb'

This differentiation solves ahead of time one of the more difficult problems for automatic parsers in English, the differentiation of prepositions from particles, as reflected in the ambiguity of (7).

7. Harry looked up the street.

H&K also differentiate the different instances of 'to', used as a preposition, above, and as a complementizer, as below:

to devote	'to'
to continue	'to'

Such syntactic pre-disambiguation of phonologically indistinguishable categories appears in other cases. For example, 'it' as subject is differentiated from 'it' as object; 'that' as conjunction is differentiated from 'that' as a relative pronoun; simple past tense verb forms ('pushed', 'sang') are differentiated from superficially identical past participle forms ('pushed', except in strong verbs, 'sung').

We have adduced the above categorization disambiguations from the four categorized sample selections which H&K present. In the categorization scheme they used (Francis & Kucera, 1979) other syntactic homophones are disambiguated in the categorization framework as well. Thus, one of the hardest problems for parsers may be solved by the categorization of the input—how to distinguish the use of a category in terms of its local syntactic function. This definitely is among the TRICS, in the sense defined above. There seems to be at least another. The categories are further differentiated

according to their frequency—frequent categories were assigned to initially more active input codes. This will tend to differentiate function from content categories, which may greatly facilitate the detection of syntactic patterns.

A more complete analysis of H&K's model awaits more complete presentation of its characteristics. We tentatively conclude that insofar as it is successful, it is because of a number of TRICS, of the same general kind as those found with simpler models without hidden units. The TRICS pre-categorize the input so that relevant grammatical regularities are directly encoded or made more accessible.

The decision about how fine-grained the categories are involves a dilemma. The extremes are to give no information in the categories, and to differentiate every possible syntactic category afforded by some grammatical theory. For example, the sample sequence (8c) could be entered either as (8a) or (8b) below:

8a. word, word, word, word, word, word

8b. definite determiner modifying the following noun, singular noun in the same phrase as the preceding determiner and both grammatical subject and thematic agent of the following verb, past-tense transitive verb, definite determiner modifying the following noun, singular noun in the same phrase as the preceding determiner and both grammatical object and thematic patient of the preceding verb, adverb modifying the preceding verb.

8c. determiner, noun, verb, determiner, noun, adverb

Regenerating sequences like (8a) is of no interest: each word boundary has a different code, which is what represents serial order in the input code. Hence, regenerating the number of words in the input would be straightforward. Learning to regenerate sequences like those in (8b) would be of little interest to Hanson & Kegl for the obverse reason: all the grammatical information we can think of is encoded within the category distinctions: if the model learned to regenerate sequences of this kind, H&K could not claim that it shows that the structural universals are not innate to the learner: in this case the structural universals are embedded in the richness of the input coding scheme, which is an 'innate' part of the model.

This contrast clarifies the dilemma which H&K face: learning on a limited categorization scheme is of limited interest, learning on a complex categorization scheme would not allow them to make claims about the induction of grammatical universals from the data. What might be of interest is a model which succeeds in regenerating input at an intermediate complexity of categorization, roughly like that which may be available to a child (see Valian, 1986). For example, if the machine can learn to regenerate sequences only stated in the basic 7–8 categories as in (8c), it might provide an existence

demonstration of the kind H&K seek. Even more impressive, would be a demonstration that it can learn to map inputs of the complexity in (8c) onto outputs of the complexity in (8b). That would be more like real learning, taking in lexically categorized input and learning to map it onto grammatically categorized output. In such a case, both the simple and complex categories could be viewed as innate. The empirical question would be: can the model learn the regularities of how to map simple onto complex grammatical categorizations of sentences? If such a model were empirically successful, it would offer the potential of testing some aspects of the empiricist hypothesis which H&K may have in mind.

Aside from TRICS, it is clear that auto association works both at the phonological and syntactic level because there are redundancies in the input: the hidden units give the system the computational power to differentiate disjunctive categories, which allows for isolation of disjunctive redundancies. We think that the models involve techniques which may be of great importance for certain applications. But, as always with artificially 'intelligent' systems, the importance of these proofs for psychological models of learning, will depend on their rate of success and the internal analysis both of the prior architecture of the model and of categories implied by the final weight patterns. Ninety percent correct after 180,000 trials is not impressive when compared to a human child, especially when one notes that the child must discover the input categories as well. In fact, it is characteristic of widely divergent linguistic theories that they usually overlap on 95 percent of the data—it is the last 5 percent which brings out deep differences. Many models can achieve 90 percent correct analysis: this is why small differences in how close to 100 percent correct a model is can be an important factor. But, most important is the analysis which the machine gives to each sentence. At the moment, it is hard for us to see how 45 associative units will render the subtlety we know to be true of syntactic structures.

8. Conclusions

We have interpreted R&M, M&K, and H&K as concluding from the success of their models that artificial intelligence systems that learn are possible without rules. At the same time we have shown that both the learning and adult-behavior models contain devices that emphasize the information which carries the rule-based representations that explain the behavior. That is, the models' limited measure of success ultimately depends on structural rules. We leave it to connectionists' productionist brethren in artificial intelligence and cognitive modelling to determine the implications of this for the algorithmic use

of rules in artificial intelligence. Our conclusion here is simply that one cannot proclaim these learning models as replacements for the acquisition of linguistic rules and constraints on rules.

9. Some general considerations

9.1. The problem of segmenting the world for stimulus/response models of learning

In the following section we set aside the issue of the acquisition of structure, and consider the status of models like R&M's as performance models of learning. That is, we now stipulate that representations of the kind we have highlighted are built into the models; we then ask, are these learning models with built-in representations plausible candidates as performance models of language acquisition? We find that they suffer from the same deficiencies as all learning models which depend on incrementally building up structures out of isolated trials.

The learning models in McClelland & Rumelhart (1986) are a complex variant on traditional—or at least Hullean—s–r connections formed in time (with secondary and tertiary connections-between-connections corresponding to between node, and hidden-node connections). The connectionist model operates 'in parallel', thereby allowing for the simultaneous establishment of complex patterns of activation and inhibition. We can view these models as composed of an extremely large number of s–r pairs, for example, the verb-learning model would have roughly 211,600 such pairs, one for each input/output Wickelfeature pair. In fact, we can imagine a set of rats each of whose tails are attached to one of the 460 Wickelfeature inputs and each of whose nose-whiskers are all attached to one of the 460 Wickelfeature outputs: on each trial, each rat is either stimulated at his tail node or not. He then may lunge at his nose node. His feedback tells him either that he was supposed to lunge or not, and he adjusts the likelihood of lunging on the next trial using formulae of the kind explored by Hull and his students, Rescorla & Wagner (1972), and others.

We will call this arrangement of rats a Massively Parallel Rodent (MPR). Clearly, the MPR as a whole can adjust its behavior. That is, if a connection between two nodes is fixed and if the relevant input and output information is unambiguously specified and reinforced, and if there is a rule for changing the strength of the connection based on the input/output/reinforcement configuration, then the model's connection strengths will change on each trial. In his review of Skinner's *Verbal Behavior* (1957), Chomsky (1959) accepted

this tautology about changes in s–r connection strengths as a function of reinforcing episodes. But he pointed out that the theory does not offer a way of segmenting which experiences count as stimuli, which as responses, and to which pairs of these a change in drive level (reinforcer) is relevant. Stimulus, response and reinforcement are all interdefined, which makes discovering the 'laws' of learning impossible in the stimulus/response framework.

We solved the corresponding practical problem for our MPR by giving it the relevant information externally—we segment which aspects of its experience are relevant to each other in stimulus/response/reinforcement triples. Accordingly, the MPR works because each rat does not have to determine what input Wickelfeature activation (or lack of it) is relevant to what output Wickelfeature, and whether positive or negative: he is hard-tailed and -nosed into a given input/output pair. Each learning trial externally specifies the effect of reinforcement on the connection between that pair. In this way, the MPR can gradually be trained to build up the behavior as composed out of differential associative strengths between different units.

Suppose we put the MPR into the field, (the verbal one), and wait for it to learn anything at all. Without constant information about when to relate reinforcement information to changes in the threshold and response weights, nothing systematic can occur. The result would appear to be as limited and circular as the apparently simpler model proposed by Skinner. It might seem that giving the MPR an input set of categories would clear this up. It does not, unless one also informs it which categories are relevant when. The hope that enough learning trials will gradually allow the MPR to weed out the irrelevancies, begs the question; which is, *how does the organism know that a given experience counts as a trial?* The much larger number of component organisms does not change the power of the single-unit machine, if nobody tells them what is important in the world and what is important to do about it. There's no solution, even in very, very large numbers of rats.

Auto-association in systems with hidden units might seem to offer a solution to the problem of segmenting the world into stimuli, responses and reinforcement relations: these models operate without separate instruction on each matching trial. Indeed, Elman & Zipser's model apparently arrives at a compact representation of input speech without being selectively reinforced. But, as we said, it will take a thorough study of the final response patterns of the hidden units to show that the model arrives at a *correct* compact representation. The same point is true of models in the style of Hanson and Kegl. And, in any case, analysis of the potential stimuli is only part of the learning problem—such models still would require identification of which analyzed units serve as stimuli in particular reinforcement relations to which responses.

So, even if we grant them a considerable amount of pre-wired architectures, and hand-tailored input, models like the MPR, and their foreseeable elaborations, are not interesting candidates as performance models for actual learning; rather, they serve, at best, as models of how constraints can be modified by experience, with all the interesting formal and behavioral work of setting the constraints done outside the model. The reply might be that every model of learning has such problems—somehow the child must learn what stimuli in the world are relevant for his language behavior responses. Clearly, the child has to have a segmentation of the world into units, which we can grant to every model including the MPR. But if we are accounting for the acquisition of knowledge and not the pairing of input/output behaviors, then there is no restriction on what counts as a relevant pair. The problem of what counts as a reinforcement for an input/output pair only exists for those systems which postulate that learning consists of increasing the probability of particular input/output pairs (and proper subsets of such pairs). The hypothesis-testing child is at liberty to think entirely in terms of confirming systems of knowledge against idiosyncratic experiences.

An example may help here. Consider two ways of learning about the game of tag: a stimulus/response model, and an hypothesis testing model. Both models can have innate characteristics which lead to the game as a possible representation. But, in a connectionist instantiation of the s/r model, the acquired representation of the game is in terms of pairs of input and output representations. These representations specify the vectors of motion for each player, and whether there is physical contact between them. We have no doubt that a properly configured connectionist model would acquire behavior similar to that observed in a group of children, in which the players successively disperse away from the player who was last in contact with the player who has a chain of dispersal that goes back to the first player of that kind. As above, the machine would be given training in possible successive vector configurations and informed about each so that it could change its activation weights in the usual connectionist way. Without that information, the model will have no way of modulating its input activation configuration to conform to the actual behavior.

Contrast this with the hypothesis testing model, also with innate structure: in this case, what is innate is a set of possible games, of which tag is one, stated in terms of rules. Consider what kind of data such a model must have: it must be given a display of tag actively in progress, in some representational language, perhaps in terms of movement vectors and locations, just like that for the connectionist model. But what it does not need is step-by-step feedback after each projection of where the players will be next. It needs instances of predictions made by the game of tag that are fulfilled: e.g., that after

contact, players' vectors tend to reverse. How many such instances it needs is a matter of the criteria for confirmation that the hypothesis tester requires. Thus, both the stimulus/response pattern association model and the hypothesis testing model require some instances of categorizable input. But only the s/r model requires massive numbers of instances *in order to construct a representation of the behavior.*

9.2. The human use for connectionist models

The only relation in connectionist models is strength of association between nodes. This makes them potential models in which to represent the formation of associations, which are (almost by definition) frequent—and, in that sense, important—phenomena of everyday life. Given a structural description of a domain and a performance mechanism, a connectionist model may provide a revealing description of the emergence of certain regularities in performance, which are not easily described by the structural description or performance mechanism alone. In this section, we explore some ways to think about the usefulness of connectionist models in integrating mental representations of structures and associations between structures.

9.2.1. Behavioral sources of overgeneralization

We turn first to the 'developmental' achievement of Rumelhart and McClelland's model of past tense learning, the overgeneralization of the 'ed' ending after having achieved better-than-chance performance on some irregular verbs. This parallels stages that children go through, although clearly not for the same reasons. R&M coerce the model into overgeneralization and regression in performance by abruptly flooding the model with regular-verb input. Given the relative univocality of the regular past tense ending, such coercion may turn out to be unnecessary—even equal numbers of regular and irregular verbs may lead to a period of overgeneralization (depending on the learning curve function) because the regular ending processes are simpler. In any case, the model is important in that it attempts to address what is a common developmental phenomenon in the mastery of rule-governed structures—at first, there is apparent mastery of a structurally defined concept and then a subsequent decrease in performance, based on a generalization from the available data. It is true that some scholars have used these periods of regression as evidence that the child is actively mastering structural rules, e.g.,

9. add 'ed' to form the past tense

Clearly, the child is making a *mistake.* But it is not necessarily a mistaken application of an actual rule of the language (note that the formula above is

not exactly a rule of the language). Rather, it can be interpreted as a performance mistake, the result of an overactive speech production algorithm which captures the behavioral generalization that almost all verbs form the past tense by adding /ed/ (see Macken, 1987, for a much more formal discussion of this type). This interpretation is further supported by the fact that children, like adults, can express awareness of the distinction between what they say, and what they know they should say (section 5.4).

There are many other examples of overgeneralization in cognitive development, in which rule-based explanations are less compelling than explanations based on performance mechanisms (Bever, 1982; Strauss, 1983). Consider, for example, the emergence of the ability to conserve numerosity judgments between small arrays of objects. Suppose we present the array on the left below to children, and ask which row has more in it (Bever, Mehler, & Epstein, 1968; Mehler & Bever, 1967). Most children believe that it is the row on the bottom. Now suppose we change the array on the left below to the one on the right, and ask children to report which row now has more: 2-year-old children characteristically get the answer correct, and always perform better than 3-year-old children.

```
* * * *        * * * *
* * * * * *     ******
```

The 3-year-olds characteristically choose the longer row—this kind of pattern occurs in many domains involving quantities of different kinds. In each case, the younger child performs on the specific task better than the older child. But the tasks are chosen so that they bring a structural principle into conflict with a perceptual algorithm. The principle is 'conservation', that if nothing is changed in the quantity of two unequal arrays, the one with more remains the one with more. The perceptual algorithm is that if an array looks larger than another, it has more in it. Such strategies are well-supported in experience, and probably remain in the adult repertoire, though better integrated than in the 3-year-old. Our present concerns make it important that the overgeneralized strategy that causes the decrease in performance is not a structural rule in any sense; it is a behavioral algorithm.

A similar contrast between linguistic structure and perceptual generalization occurs in the development of language comprehension (Bever, 1970). Consider (10a) and (10b).

10a. The horse kicked the cow
10b. The cow got kicked by the horse

At all ages between 2 and 6, children can make puppets act out the first kind of sentence. But the basis on which they do this appears to change from age

2 to age 4: at the older age, the children perform markedly worse than at the younger age on passive sentences like (10b). This, and other facts suggest that the 4-year-old child depends on a perceptual heuristic,

11. "assign an available NVN—sequence, the thematic relations, agent, predicate, patient."

This heuristic is consistent with active sentence order, but specifically contradicts passive order. The heuristic may reflect the generalization that in English sentences, agents do usually precede patients. The order strategy appears in other languages only in those cases in which there is a dominant word order. In fact, in heavily inflected languages, children seem to learn very early to depend just on the inflectional endings and to ignore word order (Slobin & Bever, 1980). The important point here is that the heuristics that emerge around 4 years of age are not reflections of the structural rules. In fact, the strategies can interfere with linguistic success of the younger child and result in a decrease in performance.

Certain heuristics draw on general knowledge, for example the heuristic that sentences should make worldly sense. 4-year-old children correctly act out sequences which are highly probable (12a), but systematically fail to act out corresponding improbable sequences like (12b).

12a. The horse ate the cookie
12b. The cookie ate the horse

The linguistic competence of the young child before such heuristics emerge is quite impressive. They perform both sentences correctly—they often acknowledge that the second is amusing, but act it out as linguistically indicated. This suggests early reliance on structural properties of the language, which is later replaced by reliance on statistically valid generalizations in the child's experience.

These examples demonstrate that regressions in cognitive performance seem to be the rule, not the exception. But the acquisition of *rules* is not what underlies the regressions. Rather, they occur as generalizations which reflect statistical properties of experience. Such systems of heuristics stand in contrast to the systematic knowledge of the structural properties of behavior and experience which children rely on before they have sufficient experience to extract the heuristics. In brief, we are arguing that R&M may be correct in the characterization of the period of overgeneralization as the result of the detection of a statistically reliable pattern (Bever, 1970). The emergence of the overgeneralization is not unambiguous evidence for the acquisition of a rule. Rather, it may reflect the emergence of a statistically supported pattern of behavior. We turn in the next section to the usefulness of connectionist models in accounting for the formation and role of such habits.

9.2.2. The nodularity of mime

Up to now, we have argued that insofar as connectionist models seem to acquire structural rules, it is because they contain representational devices which approximate aspects of relevant linguistic properties. Such probabilistic models are simply not well-suited to account for the acquisition of categorical structures. But there is another aspect of behavior to which they are more naturally suited—those behaviors which are essentially associative in nature.

Associations can make repeated behaviors more efficient, but are the source of error in novel behaviors. Accordingly, we find it significant that the most empirically impressive connectionist models have been devoted to the description of erroneous or arbitrary behavior. For example, Dell's model of speech production predicts a wide range of different kinds of slips of the tongue and the contexts in which they occur. The TRACE model of speech perception predicts interference effects between levels of word and phoneme representations. By the same token, we are willing to stipulate that an improved model of past tense learning might do a better job of modelling the mistakes which children make along the way.

All of these phenomena have something in common—they result, not from the structural constraints on the domain, but from the way in which an associative architecture responds to environmental regularities. Using a connectionist architecture may allow for a relatively simple explanation of phenomena such as habits, that seem to be parasitic on regularities in behavioral patterns. This interpretation is consistent with a view of the mind as utilizing two sorts of processes, computational and associative. The computational component represents the structure of behavior, the associative component represents the direct activation of behaviors which accumulates with practice.

Clearly, humans have knowledge of the structural form of language, levels of representation and relations between them. Yet, much of the time in both speaking and comprehension, we draw on a small number of phrase- and sentence-types. Our capacity for brute associative memory shows that we can form complex associations between symbols: it is reasonable that such associations will arise between common phrase types and the meaning relations between their constituent lexical categories. The associative network cannot explain the existence or form of phrase structures, but it can associate them efficiently, once they are defined.

This commonsense idea about everyday behavior had no explicit formal mechanism which could account for it in traditional stimulus-response theories of how associations are formed. Such models had in their arsenal single chains of associated units. The notion of multiple levels of representation and matrices was not developed. The richest attempt in this direction for

language was the later work of Osgood (1968); but he was insistent that the models should not only describe associations that govern acquired behavior, they should also account for learning structurally distinct levels of representation—which traditional s/r theories cannot consistently represent or learn (see Bever, 1968; Fodor, 1965).

There are some serious consequences of our proposal for research on language behavior. For example, associations between phrase types and semantic configurations may totally obscure the operation of more structurally sensitive processes. This concept underlay the proposal that sentence comprehension proceeds by way of 'perceptual strategies' like (12), mapping rules which express the relation between a phrase type and the semantic roles assigned to its constituents (Bever, 1970). Such strategies are non-deterministic and are not specified as a particular function of grammatical structure. The existence of such strategies is supported by the developmental regressions in comprehension reviewed above, as well as their ability to explain a variety of facts about adult sentence perception (including the difficulty of sentences like (5a) which run afoul of the NVN strategy (12)). But the formulation of a complete strategies-based parser met with great difficulty. Strategies are not 'rules', and there was no clear formalism available in which to state them so that they can apply simultaneously. These problems were part of the motivation for rejecting a strategies-based parser (Frazier, 1979), in favor of either deterministic processes (production systems of a kind; Wanner & Maratsos, 1978) and current attempts to construct parsers as direct functions of a grammar (Crain & J.D. Fodor, 1985; Ford, Bresnan, & Kaplan, 1981; Frazier, Carson, & Rayner, work in progress).

On our interpretation, connectionist methods offer a richer representational system for perceptual 'strategies' than previously available. Indeed, some investigators have suggested connectionist-like models of *all* aspects of syntactic knowledge and processing (Bates & MacWhinney, 1987; MacWhinney, 1987). We have no quarrel with their attempts to construct models which instantiate strategies of the type we have discussed (if that is what they are in fact doing); however, it seems that they go further, and argue that the strategies comprise an entire grammatical and performance theory at the same time. We must be clear to the point of tedium: a connectionist model of parsing, like the original strategies model, does not explain away mental grammatical structures—in fact, it depends on their independent existence elsewhere in the user's mental repertoire. Furthermore, frequency based behavioral heuristics do not necessarily comprise a complete theory of performance, since they represent only the associatively based components.

Smolensky (in press) has come to a view superficially similar to ours about the relation between connectionist models and rule-governed systems. He

applies the analysis to the distinction between controlled and automatic processing (Schneider, Dumais, & Schiffrin, 1984). A typical secular example is accessing the powers of the number 2. The obvious way to access a large power of two, is to multiply 2 times itself that large number of times. Most of us are condemned to this multiplication route, but computer scientists often have memorized large powers of two, because two is the basic currency of computers. Thus, many computer scientists have two ways of arriving at the 20th power of two, calculating it out, or remembering it. Smolensky suggests that this distinction reflects two kinds of knowledge, both of which can be implemented in connectionist architecture—'conscious' and 'intuitive'. 'Conscious' knowledge consists of memorized rules similar to productions. Frequent application of a conscious rule develops connection strengths between the input and the output of the rule until its effect can be arrived at without the operation of the production-style system. Thus the production-style system 'trains' the connectionist network to asymptotic performance.

The conscious production-style systems are algorithms, not structural representations. For example, the steps involved in successive multiplications by two depend on available memory, particular procedures, knowledge of the relation between different powers of the same number, and so on. Thus, this proposal is that the slow-but-sure kind of algorithms can be used to train the fast-but-probabilistic algorithms. Neither of them necessarily represents the structure. Smolensky, however, suggests further that a connectionist model's structural 'competence' can be represented in the form of what the model would do, given an infinite amount of memory/time. That is, 'ideal performance' can be taken to represent competence (note that 'harmony theory' referred to below is a variant of a connectionist system).

> It is a corollary of the way this network embodies the problem domain constraints, and the general theorems of harmony theory, that the system, when given a well-posed problem, and infinite relaxation time, will always give the correct answer. So, under that idealization, the competence of the system is described by hard constraints: Ohm's law, Kirchoff's law. It is as if it had those laws written down inside of it.

Thus, the system conforms completely to structural rules in an ideal situation. But the ability to conform to structure in an infinite time is not the same as the representation of the structure. The structure has an existence outside of the idealized performance of a model. The structure corresponds to the 'problem domain constraints'.

Smolensky directly relates the production-style algorithms to explicit potentially conscious knowledge, such as knowing how to multiply. This makes the proposal inappropriate for the representation of linguistic rules, since

they are generally not conscious, and most of them operate on abstract objects. Consider for example, the seven rules involved in the description of the pronunciation of the regular past tense. Many of them operate over abstract phones, which are by definition unpronounceable (similarly, in autosegmental terms, the structures are abstract schemata). It is not meaningful to think of explicit learning or teaching of such operations. But, this has been the problem all along: language has a rich and abstract structure, whatever the algorithms are that humans use when they speak and understand it.

9.3. The end—habits and rules

We have attempted to put the connectionist debate about rules in perspective. The debate was initially among those interested in cognitive modelling, contrasting propositionally-based algorithms such as those used in many production systems with constraint satisfaction algorithms. That form of debate has been extended to include the question of whether connectionist models can learn to exhibit rule-governed behavior without acquiring rules. With differing shades of emphasis, there have been three answers to the specific challenge to rule-based theories of language acquisition posed by R&M's and H&K's models.

- Some mental processes require rules anyway (Fodor & Pylyshyn)
- R&M's model, in particular, does not work, empirically or theoretically (Pinker & Prince)
- Our emphasis in this paper has been on the fact that the connectionist models we have considered arrive at rule-like regularities in language behavior only insofar as the models already contain architectures and devices explained in humans by mental representations of categorical rules.

This is not surprising since there is no natural way for an associative device to represent categorical rules: hence, as adult models, such artificially intelligent systems require a built-in sensitivity to the rules or their equivalents. As acquisition devices, the connectionist machines share the limitation of all models which change their behavior as the accumulated result of pairing stimuli and responses: namely, they can never discover structural rules. The positive feature of connectionist models is that they provide a rich associative framework for the description of the formation of complex habits.

In sum, we have reminded the reader that the structure of human behaviors such as language cannot be explained by associative networks. But it is equally obvious that *some* behaviors are habits—the result of associations

between structurally defined representations. We think that connectionist models are worth exploring as potential explanations of those behaviors: at the very least, such investigations will give a clearer definition of those aspects of knowledge and performance which cannot be accounted for by testable and computationally powerful associationistic mechanisms.

References

Anderson, J.A. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Bates, E., & MacWhinney, B. (1987). Competition variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bever, T. (1968). A formal limitation of associationism. In T. Dixon & D. Horton (Eds.), *Verbal behavior and general behavior theory.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bever, T. (1970). The cognitive basis for linguistic universals. In J.R. Hayes (Ed.), *Cognition and the development of language* (pp. 277–360). New York, NY: Wiley & Sons, Inc.

Bever, T. (1975). Psychologically real grammar emerges because of its role in language acquisition. In D. Dato (Ed.), *Developmental psycholinguistics: Theory and applications* (pp. 63–75). Georgetown University Round Table on Languages and Linguistics.

Bever, T. (1982). Regression in the service of development. In T. Bever (Ed.), *Regression in mental development* (pp. 153–188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bever, T., & Langendoen, D. (1963). (a) The formal justification and descriptive role of variables in phonology. (b) The description of the Indo-European E/O ablaut. (c) The E/O ablaut in Old English. *Quarterly Progress Report RLE.* MIT, Summer.

Bever, T., Carroll, J., & Miller L.A. (1984). Introduction. In T. Bever, J. Carroll & L.A. Miller (Eds.), *Talking minds: The study of language in the cognitive sciences.* Cambridge, MA: MIT Press.

Bever, T., Mehler, J., & Epstein, J. (1968). What children do in spite of what they know. *Science, 162,* 921–924.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Bybee, J., & Slobin, D. (1982). Rules and schemes in the development and use of the English past tense. *Language, 58,* 265–289.

Chomsky, N. (1959). Review of Skinner's Verbal Behavior. *Language, 35,* 26–58.

Chomsky, N. (1964). The logical basis of linguistic theory. In Proceedings of the 9th International Conference on Linguistics.

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York, NY: Harper and Row.

Crain, S., & Fodor, J.D. (1985). How can grammars help parsers? In D.R. Dowty, L. Kartunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational and theoretical perspectives.* Cambridge: Cambridge University Press.

Dell, G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 3, 283–321.

Elman, J.L., & McClelland, J.L. (1986). Exploiting the lawful variability in the speech wave. In J.S. Perkell & D.H. Klatt (Eds.), *Invariance and variability of speech processes.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J.L., & Zipser, D. (1987). Learning the hidden structure of speech. USCD Institute for Cognitive Science Report 8701.

Feldman, J. (1986). Neural representation of conceptual knowledge. University of Rochester Cognitive Science Technical Report URCS-33.

Feldman, J., & Ballard, D. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205–254.

Feldman, J., Ballard, D., Brown, C., & Dell, G. (1985). Rochester Connectionist Papers: 1979–1985. University of Rochester Computer Science Technical Report TR-172.

Fodor, J.A. (1965). Could meaning be an r_m? *Journal of Verbal Learning and Verbal Behavior, 4*, 73–81.

Fodor, J.A., Bever, T.G., & Garrett, M. (1974). *Psychology of language.* New York: McGraw Hill.

Fodor, J.A., & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71, this issue.

Ford, M., Bresnan, J., & Kaplan, R. (1981). A competence-based theory of syntactic closure. In J. Bresnan (Ed.), *The mental representation of grammatical relations.* Cambridge, MA: MIT Press.

Francis, W.N., & Kucera, H. (1979). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers.* Providence, RI, Department of Linguistics, Brown University.

Frazier, L. (1979). On comprehending sentences: Syntactic parsing strategies. Doctoral Thesis, University of Massachusetts.

Frazier, L., Carson, M., & Rayner, K. (1985). Parameterizing the language processing system: Branching patterns within and across languages. Unpublished.

Garrett, M. (1975). The analysis of sentence production. In G. Bower (Ed.), *The psychology of learning and motivation.* New York: Academic Press (pp. 133–177).

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis, 2*, No. 1.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), *Syntax and semantics 3: Speech acts.* New York, NY: Seminar Press.

Grossberg, S. (1987). Competitive learning from interactive activation to adaptive resonance. *Cognitive Science, 11*, 23–63.

Halle, M. (1962). Phonology in generative grammar. *Word, 18*, 54–72.

Hanson, S.J., & Kegl, J. (1987). PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. *Proceedings of the Ninth Annual Cognitive Science Society Meeting.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hinton, G., & Anderson, S. (Eds.) (1981). *Parallel models of associative memory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hinton, G., & Sejnowski, R. (1983). Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C.

Hinton, G., & Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA: Vol. 79. Biophysics* (pp. 2551–2558).

Kuczaj, S. (1978). Children's judgments of grammatical and ungrammatical irregular past-tense verbs. *Child Development, 49*, 319–326.

Kuroda, S.Y. (1987). Where is Chomsky's bottleneck? Reports of the Center for Research in Language, San Diego, Vol. 1, No. 5.

Langackre, R. (1987). The cognitive perspective. In Reports of the Center for Research in Language, San Diego, Vol. 1, No. 3.

Macken, M. (1987). Representation rules and overgeneralizations in phonology. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 367–397). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1987). The competition model of the acquisition of syntax. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249–308). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision.* San Francisco, CA: Freeman.

McClelland, J., & Elman, J. (1986). Interactive processes in speech perception: The TRACE model. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure*

of cognition: Vol. 2. Psychological and biological models. Cambridge, MA: MIT Press.

McClelland, J. & Kawamoto, A. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models.* Cambridge, MA: MIT Press.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: Part 1: An account of basic findings. *Psychological Review, 88*, 5, 60–94.

McClelland, J. & Rumelhart, D. (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models.* Cambridge, MA: MIT Press.

Mehler, J., & Bever, T. (1967). A cognitive capacity of young children. *Science*, Oct. 6, 141.

Miller, G.A. (1962). Some psychological studies of grammar. *American Psychologists, 17*, 748–762.

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

Neches, R., Langley, P., & Klahr, D. (1987). Learning, development and production systems. In D. Klahr, P. Langley & R. Neches (Eds.), *Production system models of learning and development.* Cambridge, MA: MIT Press.

Osgood, C.E. (1968). Toward a wedding of insufficiencies. In T. Dickson & D. Horton, *Verbal behavior and general behavior theory*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

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

Reich, P.A. (1969). The finiteness of natural language. *Language, 45*, 831–843.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokosy (Eds.), *Classical conditioning.* New York: Appleton-Century-Crofts, pp. 64–99.

Rosenblatt, F. (1962). *Principles of neurodynamics.* New York, NY: Spartan.

Rumelhart, D., Hinton, G.E., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland & the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Formulations.* Cambridge, MA: MIT Press/Bradford Books.

Rumelhart, D., & McClelland, J. (1982). An interactive activation model of context effects in letter perception: Part 2: The contextual enhancement effect and some tests of the model. *Psychological Review, 89*, 1, 60–94.

Rumelhart, D., & McClelland, J. (Eds.) (1986a). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D., & McClelland, J. (1986b). On learning the past tenses of English verbs. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models.* Cambridge, MA: MIT Press.

Sadock, J. (1974). *Toward a linguistic theory of speech acts.* New York: Academic Press.

Sampson, G. (1987). A turning point in linguistics: Review of D. Rumelhart, J. McClelland and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Times Literary Supplement*, June 12, 1987, p. 643.

Sapir, E. (1921–49). *Language.* New York: Harcourt, Brace and World.

Savin, H., & Bever, T. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior, 9*, 295–302.

Schneider, W., Dumais, S., & Shriffrin, R. (1984). Automatic and control processing and attention. In R. Parasaraman & D. Davies (Eds.), *Varieties of attention.* New York: Academic Press, Inc.

Skinner, B. (1957). *Verbal behavior.* New York, NY: Appleton-Century-Crofts.

Slobin, D. (1978). A case study of early language awareness. In A. Sinclair, R. Jarvella, & W. Levelt (Eds.), *The child's conception of language.* New York, NY: Springer-Verlag.

Slobin, D., & Bever, T.G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order. *Cognition, 12*, 219–277.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Smolensky, P. (in press). The proper treatment of connectionism. *Behavioral and Brain Sciences.*

Strauss, S., & Stavy, R. (1981). U-shaped behavioral growth: Implications for theories of development. In W.W. Hartup (Ed.), *Review of child development research* (Volume 6), Chicago: University of Chicago Press.

Tanenhaus, M.K., Carlson, G., & Seidenberg, M. (1985). Do listeners compute linguistic representations? In D. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 459–408). Cambridge: Cambridge University Press.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan and G. Miller (Eds.), *Linguistic theory and psychological reality.* Cambridge, MA: MIT Press.

Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76,* 1–15.

Résumé

Les modèles connexionnistes des conduites linguistiques acquises qui ont été récemment proposés, incorporent en fait des représentations linguistiques fondées sur un système de règles. Des modèles connexionnistes similaires pour l'acquisition du langage possèdent de même un appareillage et une architecture ad hoc qui leur fait mimer les effets des règles. Les modèles connexionnistes en général ne sont pas adéquats pour rendre compte de l'acquisition de connaissances structurelles: ils nécessitent des structures prédéterminées, même pour simuler des faits linguistiques élémentaires. De tels modèles sont plus appropriés pour décrire la formation d'associations complexes entre des structures qui sont déjà représentées de manière indépendante. Ceci rend les modèles connexionnistes des outils potentiellement importants pour étudier les relations entre des conduites fréquentes et les structures qui sous-tendent les connaissances et les représentations. Ces modèles peuvent offrir de puissants moyens informatiques pour démontrer les limites d'une description associationiste des conduites.