Connectionist Approaches to Natural Language Processing

edited by

Ronan G. Reilly

Department of Computer Science, University College Dublin, Ireland

Noel E. Sharkey

Department of Computer Science, University of Exeter, U.K.

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Hove (UK) Hillsdale (USA)

Lawrence Erlbaum Associates Ltd., Publishers
27 Palmeira Mansions
Church Road
Hove
East Sussex, BN3 2FA
U.K.

British Library cataloguing in Publication Data

Connectionist approaches to natural language processing
I. Reilly, Ronan G., *1955–* II. Sharkey, N.E. (Noel E)
006.35

ISBN 0-86377-179-3

Typeset by DP Photosetting, Aylesbury, Bucks.
Printed and bound by BPCC Wheatons, Exeter.

8 The Demons and the Beast—Modular and Nodular Kinds of Knowledge

Thomas G. Bever
The University of Rochester, Wilson Blvd., Rochester, NY 14627, U.S.A.

Chaos often breeds life, when order breeds habit
—Henry Adams

The human brain has to develop analogies to make up for its limitations
—Alan Turing

INTRODUCTION AND SUMMARY

An enduring insight of behavioural science is that most actual behaviour is a function of habit. This was the basis for the enduring popularity of S-R psychology, despite the many empirical inadequacies that ultimately lead to its downfall in the "cognitive revolution." Connectionism appears to be rehabilitating the image of associationistic models of behaviour, by increasing the adequacy of associationistic models to account for structural phenomena invoked by cognitive psychologists as the basis for rejecting S-R psychology.

The negative thesis of this chapter is that connectionist models can capture only habits and are therefore inadequate in principle to capture structural processes. The positive thesis of this chapter is that connectionist models can capture only habits, and therefore are an important new tool in the study of how habitual knowledge interacts with structural processes.

The first part of the chapter reviews evidence that many behaviours are the result of a complex interaction of structural and habitual processes. The second part reviews several recent important connectionist models of a large segment of English knowledge, primarily the verb-learning model of Rumelhart and McClelland, and the word-reading model of Seidenberg and McClelland: In each case, careful examination of the model shows that it has structural

properties built into its representation and feeding schemes—thus, the models neither discover nor circumvent knowledge structures; rather, they reflect the relevant probabilities in their input, given their particular representational schemes. The third part presents a framework for modelling language acquisition, in which the child is seen as building up an interaction between structural and associative representations of language. A connectionist model, Baby Clauseau, demonstrates how connectionist models can be utilised to study associatively learnable patterns in the environment. Baby Clauseau recognises only 100 frequent words of motherese and where actual utterances end: It learns, from actual motherese discourses, to segment utterances into distinct phrases which are linguistically appropriate. This exemplifies the positive role of connectionism in the study of how habits and structures may interact.

STRUCTURE AND HABITS—THE KNOWLEDGE AND THE POWER

Small habits well pursued betimes
May reach the dignity of crimes.

—Hannah More

One of the outstanding facts about behaviour is that most of the time we do what we do most of the time. Consequently, much of what psychologists study is based on accumulated habits; patterns of behaviour that are well-oiled. This makes psychology hard to do, if you think that there might be more underlying behaviour than probabilistic habits alone. Reasoning and language are among those skills suggesting that the mind is a non-probabilistic computational machine that carries out discrete operations on symbols. For example, if you see the array on the left in (1) transformed into the one on the right by compression of the upper row, you will reason that the upper row still has more circles than the lower does. Your ability to do this involves manipulation of symbols at an abstract level of representation. It runs something like this: Nothing was added or taken away when the row was compressed, so the number of dots in the two arrays must still be the same. There is nothing concrete or probabilistic about this judgement; it is symbolic and categorical.

1. o o o o o o o o o o o o o oooooooooooo
   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Similarly, if you know English you can decide that (2b) is licensed by (2a) but (2d) is not licensed by (2c). Here, too, your judgement is categorical, following the differentiation of lexical passive adjectives (e.g. "unsuspected") and verbal passives.

2. a. The girl was surprised.
   b. Somebody (or something) surprised the girl.
   c. The girl was unsurprised.
   d. *Somebody (or something) unsurprised the girl.

Even simple observations like these necessitate the development of computational models in which symbols represent categories, and discrete processes state the relations between the symbols. Mental theories of number and grammars of language are typical examples of such models. The independent effect of structural knowledge is hard to bring out in adults because the pathways of ordinary behaviour are so well-practiced: For example, if you simply looked at the array on the right in (1), you might suspend your reasoning capacity and decide that the short line of "o"s is, in fact, less numerous than the line of "0"s, simply because it looks that way—in general it is the case that shorter arrays have fewer components. Similarly, the way one actually understands sentences may draw on probabilistic properties of languages. For example, one may comprehend the relation between the subject and verb in (3a) based on the generalisation that, unless, otherwise marked, as in (3b) and (3c), agents precede their verbs.

3. a. The girl hit the wall.
   b. It's the wall that the girl hit.
   c. The wall was hit by the girl.

These simple observations (confirmed in a variety of experimental settings; see Bever, 1970; 1975b, for reviews) justify the distinction between two kinds of mental entities—a structural demon and an habitual beast. By definition, in everyday behaviour, the habitual beast overwhelms the structural demon. Most number-related behaviour does not involve rendering judgement about unusual configurations, and most language behaviour does not involve rendering grammaticality intuitions about sentences in the abstract: Most of the time, we rely on habitual appearances to make judgement of quantity, and use habitual processes to understand sentences. Accordingly, it is extremely difficult to demonstrate the effect of structural processes in adult behaviour. Indeed, it is a real question whether structural processes play a determining role in everyday behaviour at all, once they are surrounded by a habitual overlay.

A logical solution to this empirical question is also an empirical one—to examine what young children do before they have enough experience to acquire generalisations. Often, a child's behaviour can display the formation of behavioural systems, drawn out in time in such a way that we can disentangle categorical mental structures from habitual overlays on them. On this view, one expects that children will first display categorical and structural processes in relatively pure form, and will subsequently suffuse them with habitual processes based on the statistics of their experience.

However, for a time, various classical demonstrations were taken to show that children actually start out with behavioural strategies and work from them into structural representations and processes. For example, the transformation from the left to the right array in (1) is a variant on a classical Piagetian paradigm, the study of the principle of "conservation". The typical finding is that children do not master the correct answer until they are about six years old—at age four, they seem dependent on the surface appearance only to make their judgement of quantity: They clearly rely on the length strategy to make numerical judgement. Similarly, children at age four systematically misunderstood passive sentences, and object-first sentences like (3b) and (3c). This suggests that they are using the first-noun = agent strategy as the basis for understanding simple sentences. These lines of research supported the view that the child starts out basing its behaviour on statistical generalisations, and subsequently develops structural representations.

The logic of how that might work was never clear: How does one arrive at categorical representations from statistical generalities? One cannot. Equally problematic is the fact that my investigations of two-year-olds (with Jacques Mehler; see Bever, 1982, for a review) suggested that the four-year-olds' reliance on statistical generalisations arises out of more basic structural capacities. For example, two-year-old children characteristically perform conservation tasks like (1) well above chance; similarly, they understand object-first cleft sentences like (3b) quite well. That is, the four-year-old's behaviour represents a change from a dependence on structural representations to a dependence on statistical generalisations. The typical performance curve on the unusual kinds of tasks is U-shaped, which has lead to the view that the child's behaviour shows a "regression".[1]

The observed distinction between children's early categorical capacities and their eventual dependence on generalisations showed that such generalisations could replace structural processing in many instances of comprehension. This was part of the motivation for a strategy-based theory of sentence comprehension in adults (Bever, 1970). On the strategy-based view of comprehension, adults utilise a set of perceptual strategies that represent simultaneous constraints on mapping surface sequences onto underlying semantic representa-

[1]There have been two classes of proposals about the function of behavioural regressions: on one view, the statistical generalisations extend the application of initially limited structural capacities (Bever, Mehler, & Epstein, 1968); on the second view, the apparent regressions represent a shift from one kind of representation to another (Bowerman, 1982; Karmiloff-Smith, 1986; Langer, 1982). Either or both of these views might turn out to be right in the end, but they are hopelessly post hoc. For example, there are other ways to correct the limitations on initially limited structural capacities, most notably to develop the adult form of the capacity; similarly, there is no general explanation for the shift from one kind of representation to another.

tions, e.g. (4). Each of these strategies may be statistically supported, though each is subject to contravention in specific cases.

4. a. N – V – N = Agent Action Object.
   b. Animate nouns are agents.
   c. BE Verb + pastparticiple BY . . . indicates an exception to (a).

The perceptual strategies theory of comprehension was also motivated by the apparent failure to show how grammatical knowledge was embedded directly in language behaviour. In the first heyday of transformational psycholinguistics, it was thought that grammatical rules corresponded to mental operations. This underlays the "derivational theory of complexity," the theory that the behavioural complexity of a sentence corresponds to the number of grammatical rules in its derivation. Careful experimental research finally invalidated the derivational theory of sentence complexity, which at least temporarily destroyed the view that the grammar is directly related to the comprehension mechanism (Fodor, Bever, & Garrett, 1974). By the early 70s, the received word was that there is psychological evidence for abstract linguistic representations, but not for the computational rules which map one level of representation onto another. Perceptual strategies were involved as the probalistically valid processes that arrived at linguistic representations without grammatical computation.

The strategies-based theory of sentence comprehension did not spark a great deal of research, for several reasons. First, it is very difficult to ascertain which statistical properties of sentences are reliable cues—an extensive construction count would have been required to assess the frequency with which particular kinds of sequences correspond to particular phrased and semantic relations. In other words, it became necessary to assess the *cue validity* of surface forms for underlying representations, in the Brunswikian (1956) sense. The second difficulty was taken to be more telling: the strategies-based comprehension model is not computational. The strategies apply simultaneously as constraints on the mapping relations between outer form and semantic analysis, but do not specify how the comprehension mapping is carried out. Finally, it did not explain how comprehension works when the strategies fail. For example, most of the strategies are inconsistent with (5a), and most of them are consistent with the initial part of (5b) in a misleading way; yet these sentences can be understood. Hence, the comprehension system had to include either a set of apparently limitless backup strategies or a way of accessing linguistic knowledge as a last resort.

5. a. The girls were unimpressed by midnight
   b. The horse raced past the barn fell

The result of these considerations is that the strategies-based comprehension

system was extrapolated only by those who denied the existence of grammatical structures altogether (e.g. Bates & MacWhinney, 1987). Those who accepted the evidence for a categorical representation of language rejected strategies as vague, unformulated, and necessarily incomplete (e.g. Frazier, 1979).

Current connectionist models in artificial intelligence seem to offer a way in which one can meet the difficulties with the original formulation of the perceptual strategies model (for current reviews, see Feldman & Ballard, 1982; Hinton & Sejnowski, 1986; Rumelhart, Hinton, & Williams, 1986). In the connectionist framework, output behaviour is defined in terms of nodes that are active and inactive: Each node itself is activated by a network of connections from a set of input nodes. One result of this kind of modelling is that activation of different input nodes can be applied simultaneously to the same set of output nodes. For (a toy) example, we can envisage a set of input nodes which categorise each phrase in a linguistic sequence on such dimensions as animacy, surface order, and so on. These nodes could be mapped onto an output set, which represents the semantic function with which the phrases are paired. Clearly, the activation strength from an input "animacy" node to an output "agent" node would be high, as would the strength from the input "first nounphrase" node; conversely, the connection from "inanimate" and "second nounphrase" would be stronger to the semantic "patient" node. When given a particular input, all these connections are activated simultaneously, so the output is effectively the average of the connection strengths from the input nodes. In this way, the different constraints can apply simultaneously, as envisaged in the strategies-based model of speech perception.

Such models can also isolate the statistical regularities in the input/output relations. They can be trained by giving them correct input/output sets of nodes, and by adjusting the strength of the connections between nodes on each training trial: Whatever regularities occur in the input/output pairs can gradually exert themselves in the form of differentiated connection strengths. To continue the linguistic example, one could imagine presenting a model with data pairs consisting of a description of the input and an output description of the semantic relations assigned to each phrase in it. On each such presentation, the input/output connections are adjusted in such a way as to increase incrementally the likelihood that the output would occur, given the input. With enough pairings to represent the probabilistic facts that are true of such input/output pairings, the model will reflect them in the accumulated pattern of connection strengths. (For examples of just such toy models, see McClelland and Kawamoto [1986] and St. John and McClelland [1990]).

Connectionist models seem to offer new hope for the strategies-based model of comprehension. Indeed, they might provide a general framework for understanding the relation between the structural demon and the statistical beast in a wide range of behavioural systems. Such a model could offer a third

explanation for the formation of statistically based behavioural generalisations: They automatically arise when any structural capacity is embedded in an otherwise uncommitted system. On this view, each innate structural mental mechanism is situated initially in a sea of uncommitted units: As the mechanism performs its computational work, transforming one symbolic representation into another, the uncommitted units inevitably form direct associative connections between the different representations defined by the computational mechanism. For example, a simple grammar may specify for the child that one of the basic options for word order is that subjects precede objects. Once the child has determined that this property is true of English, then it can understand active sentence orders. From this experience, a generalisation is possible: The first phrase in a proposition is always the agent (4a). The computational mechanism for English does not specify this, but it is an inevitable generalisation out of the child's actual capacities, and the statistical properties of the sentences it experiences. Hence, the generalisations are without direct causes, but arise automatically from the interaction of internally specified representations and environmental information.

A major problem with the strategies-based model was that it did not specify the relation between the grammar and the strategies. The combination of a connectionist and structural component might offer an explanation of what has been a riddle: How can there be behavioural evidence for abstract levels of representation in comprehension, but no direct evidence for the computational rules which interrelate and thereby define those levels? On the hybrid model I have in mind, the representations are defined by the child's computational system for language (its grammar); probability-based pairings of representations at different levels automatically emerge, and become the active basis for mapping outer representations into inner ones during comprehension. Hence the representations are real, but the complex computational processes that define them are replaced by efficient probabilistic processes associating the representations.

This seems to be an extremely attractive solution to the problem of accounting for the effect of the frequency beast on the computational demon. There are numerous specific models that we can examine to see if they reveal an interaction of associative and structural knowledge. This discussion is limited to some of those models, which are alleged to learn something about a large subset of natural language. In the best instances, the models suggest that the beast/demon formulation of cognitive modelling may be a viable one—those connectionist models which seem to work by association actually *presuppose* the structural representations defined in grammars. In the specific instances, each connectionist model manages to sneak enough sensitivity to the relevant linguistic structure to guarantee that it will converge on linguistically sensitive behaviour, once it is trained with cases exhibiting the linguistic constraints.

A MODEL THAT LEARNS SOME MORPHOLOGY

Our life is like some vast lake that is slowly filling with the stream of our years. As the waters creep surely upward the landmarks of the past are one by one submerged. But there shall always be memory to lift its head above the tide until the lake is overflowing.

—Bisson

The currently most notorious model is one that purports to learn to produce the past tense of English verbs. Verbs come in two flavours; *regular* (add -ed to form the past) and *irregular* (typically, change vowel colour, in a system partially derived from the Indo-European ablaut).Rumelhart and McClelland (1986; R&M) implemented a connectionist model that learns to associate past tense with the present tense of both the regular and irregular 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 neighbourhood. For example, "bet" would be represented as composed of the following Wickelphones:

6. 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 (6), 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 suffix of the preceding unit. That rule leads to only one output representation for the three Wickelphones in (6), namely "b . . . e . . . t." R&M assign a set of distinctive phonemic features to each phone within a Wickelphone. There are four feature dimensions, two with two values and two with three, yielding ten individual feature values (see Table 8.1). This allows them to represent Wickelphones in feature matrices: For example the central /e/ in "bet" would be represented as shown in (7):

7. Dimension 1 Vowel
   Dimension 2 Low

Dimension 3	Short
Dimension 4	Front

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 suffix phone.

8.

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

There are about 1000 potential Wickelfeatures of this kind (10 prefix values × 10 central phone values × 10 suffix values).

Wickelfeature representations of the words occur at the input and the output with 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 Fig. 8.1. 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

TABLE 8.1
Categorisation 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	^	U	u
Low	A	e	I	a/ɑ	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*; ^ = 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*; ɑ = a in *father*; W = ow in *cow*; * = aw in *saw*; o = o in *hot*.

From Rumelhart & McClelland (1986, p. 235, their Table 2).

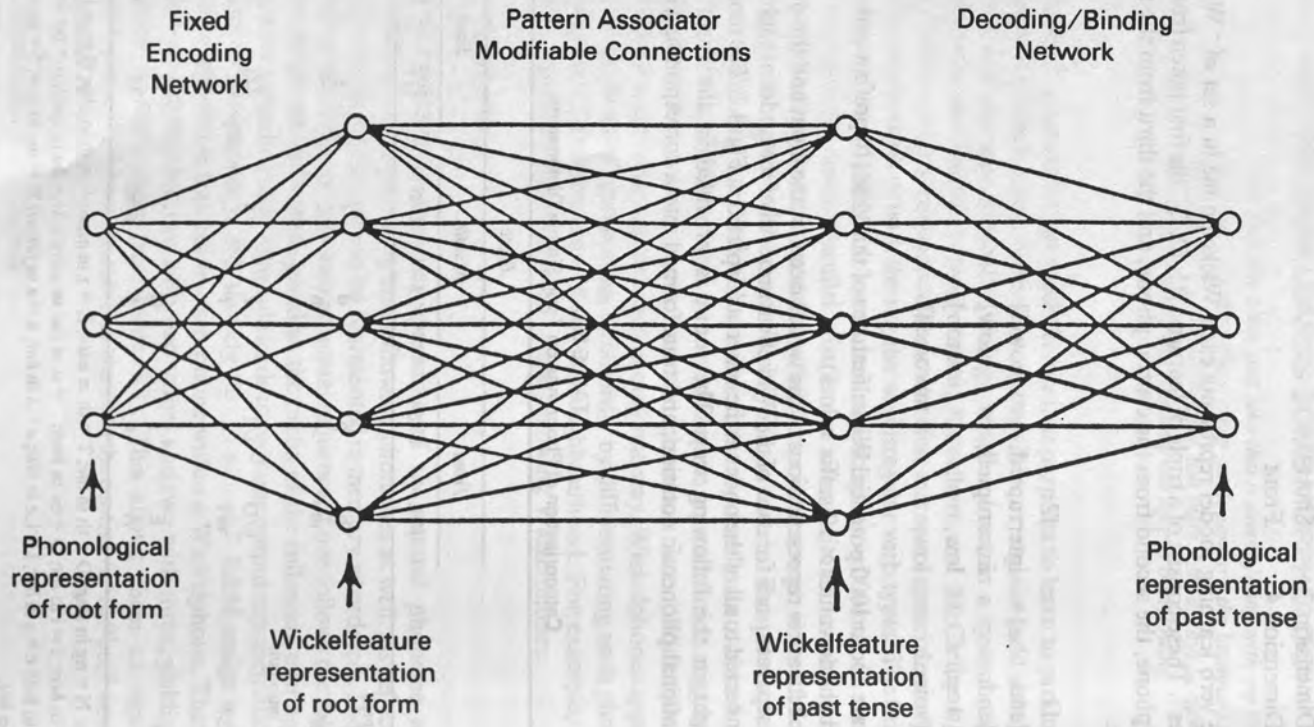


FIG. 8.1. The basic structure at the verb learning model. (From Rumelhart & McClelland (1986b), p. 222, their Figure 1).

Wickelfeatures. The input layer of Wickelfeatures is activated by the input. Each input node is connected to each output node with a specific connection 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 nodes to which it is connected.

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 the converse. The model 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 that were incorrect on the trial: Lower the weight and raise the threshold for all the output nodes that were incorrectly activated; do the opposite for nodes that were incorrectly inactivated.

The model 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 associate their past tense correctly in most cases. Hence, the model appears to learn, given a finite input, how to generalise to new cases. Furthermore, the model appears to go through several stages of acquisition; these correspond to the stages of learning the past tense of verbs that 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) "overgeneralise" the attachment of the majority past form, –ed and its variants, so that they 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 behaviours 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 behaviour. That is, this model is a potential demonstration that rule-governed behaviour 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 general, because of this model in particular (Langackre, 1987; Sampson, 1987).

The model, however, is not without critics. Fodor and Pylyshyn (1988) argued that such models cannot work in principle; Pinker and Prince (1988) argued that this model did not, in fact, work. Lachter and Bever (1988) took a different approach: We stipulated for purposes of argument that the model works, and then investigated *why* it does so. Our forensic method was the following. We examined each arbitrary feature of the model with the rule system in mind, and

asked: Would this facilitate or inhibit the behavioural emergence of data which looks like that governed by the past tense rules? Without exception, we found that each "arbitrary" decision would facilitate the emergence of such behaviour. That is, we found that a set of apparently benign simplification devices set the model to be sensitive to the rule-governed properties of the input.

In the formation of the past tense, there are two important features of the present verb form: the final phone (for regular past tenses), and the medial vocalic segment (for irregular past tenses). We described a number of representational devices of the model that insured it would be sensitive to these linguistic properties of past tense formation. (Lachter and Bever acronymically referred to those devices as TRICS—The Representations It Crucially Supposes).

The first simplification of the model involves reducing the number of within-word Wickelfeatures from about 1000 to 260. R&M did this but in an idiosyncratic way: They required that all Wickelfeatures have the same dimension for f1 and f3, whereas f2 can range fully across all feature dimensions and values. Accordingly, the potential Wickelfeatures for the vowel /e/ in "bet" in (9a) are possible; those in (9b) are not.

9. a. [interrupted, vowel, interrupted]
[voiced, mid, unvoiced]
[front, short, middle]
b. [interrupted, vowel, stop]
[stop, vowel, unvoiced]
[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 heightens the relative informativeness f2, since f1 and f3 are mutually predictable, but not with f2. This restriction is entirely arbitrary from the standpoint of the model; but it is an entirely sensible move if the goal were to accommodate to a rule-based account of the phenomena in which the relevant information in a Wickelphoneme is actually in f2. The use of centrally informative Wickelphones automatically emphasises f2.

R&M 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 in (10) are among the possible features for the /et#/ in /bet/.

10. [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 recognisable. That is, the Wickelfeature representation of the words exhibits the property of "boundary sharpening".

In order to generalise to new cases, these models must have a way of blurring the input so that the model learns on imperfect input. R&M did this by allowing input nodes to be activated if either f1 or f3 were incorrect, so long as f2 is correct. That is, the fidelity of the relationship between an input Wickelphone and the nodes that are actually activated is subject to "peripheral blurring". This effect is not small: The blurring was set to occur with a probability of 0.9 (this value was determined by R&M after some trial and error with other values). That is, a given Wickelnode is activated 90% of the time when either the input does not correspond to its f1 or f3. It can always count on f2, however. 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 earlier. 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. It also explains how the model generalises to new cases—by *failing to discriminate them* in the input of old cases. This is a charming resuscitation of the old psychological saw—that generalisation *is* the failure of discrimination—surely, recidivist behaviourism when applied to language.

The period of overgeneralisation of the regular past at the 11th cycle of trials also depends on a real trick. 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% are regular. Thus, on the 11th cycle alone, the model is given more instances of regular verbs than in all the training trials it has received before. It is no wonder, therefore, that the regular past ending immediately swamps the previously acquired irregulars. R&M defend this arbitrary move by suggesting that children also experience a sudden surge of regular past tense experience. There are no acquisition data showing anything of the sort (see Pinker & Prince, 1988, who compile evidence to the contrary). Furthermore, if there were to be 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 (with a constant lexical memory) 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 over-

generalising, it would be very difficult to decide which was the cause and which the effect.

It is the spirit of connectionist learning models to represent how people learn by inference from masses of individual experiences. Extracting the frequency variation in linguistic experience is what *linguists* do—they are interested in possible patterns, not in the frequency with which individual instances are experienced. Yet R&M presented each verb the same number of times, despite wide variation in actual frequency, thus predigesting the input for their model in much the same way a linguist does—by ignoring real frequency information. This is probably the most important trick of all—and it is absolutely clear why they did it. Irregular past tense verbs are by far and away the most frequently occurring tokens. Hence, if R&M had presented their model with data corresponding to the real frequency of occurrence of the verbs, the model would have learned all the irregulars, and might never receive enough relative data about regulars to learn them. One cannot fault R&M as computer engineers for simplifying the input in this way, but it vitiates any claim that this is a plausible inference-based learning model, and it clarifies further why it is the linguistically relevant patterns that the model tends to extract.

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. It seems obvious 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 an unnaturally restricted vocabulary, see Pinker & Prince, 1988; Savin & Bever, 1970). But Wickelphones also give equal weight to the preceding and following phone, whereas it is the central phone that 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 sub-phones 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 that happen when morphemes are adjoined. Finally, the behavioural learning properties of the model were ensured by making the model learn slowly, levelling all actual frequency information 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 behavioural regularities described in rule-governed accounts, but without any rules. Pinker and Prince (1988) demonstrate that, in fact, the model is not adequate, even for the basic facts: Hence, the first claim for the model is not correct. Lachter and Bever (1988) showed further, that, even if the model *were* empirically adequate, it would be because the model's architecture and method of data predigestion is designed to extract rule-based regularities in the input data. The impact of the rules for past-tense learning is indirectly embedded in the

form of representation and the treatment of the input data: Even Wickelfeatures involve a linguistic theory with acoustic segments and phonological features within them; the work of the special representational devices to render available the segmental phoneme, and emphasise boundary phonemes in terms of segmental features. That is, garbage in/garbage out: Regularities in/regularities out.

One might argue that the special devices are a theory of the child's morphophonological mind—that is, they constitute the theory of what underlies phonological universals. To check for that, one must examine the implications of each of the TRICS for phonological universals. None of the special devices fares well under this kind of scrutiny. For example, "central-informativeness" makes it hard to learn processes in which f1 and f3 are marked for different dimensions. Since such processes are common, the universal predictions this makes are incorrect. Similarly, sharpening information at word boundaries 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 affect segments at the boundaries of morphemes when they are combined. Unfortunately, such processes do not seem to predominate cross-linguistically over within-word processes.

One of the most unrealistic aspects of the Wickelphonological feature system is that a single feature is represented by only one node, regardless of how many times it appears in a given word. But it is often the case that a word can have more than one instance of the same feature: For example, in "deeded", almost all the features with f2 centred on the vowels are identical and thus make the word simpler to represent in Wickelnodes than "seeded", by about 12 Wickelfeatures—this means that learning the past tense of "deed" should be easier than for "seed". The particular configuration of features in R&M makes feature repetitions possible even in monosyllabic words. For example, the number of separate Wickelfeatures for the internal sounds in "fazed" is 55% larger than the corresponding number of separate Wickelfeatures for "dozed". There does not, however, seem to be any evidence that learning the past for one is easier than for the other.

Finally, the child is exposed to tokens, not types: It is difficult to see how the child could learn if it ignored the past-tense experience of a verb until all the other verb past-tense forms have been attended to at least once. One can *imagine* a model which does this—i.e. a model with a "McRumelwell's Demon", which rejects all verbs from the inner sanctum of weight-change until all the other verbs have been admitted once. But this is a demon with considerable analytic powers; knowledge about past forms of verbs that have appeared, and about which ones are being awaited. Furthermore, the demon knows just when to expand the list of to-be-learned past-tense forms (the 11th cycle). All of this without a word-level representation of words in Wickelphones!

So, in the model, the representations it crucially assumes do not define a plausible set of linguistic universals. (As Lachter and Bever [1988, p. 213] put it, "Trics aren't for kids.")[2]

EVIDENCE FOR NODES UNSEEN—SOME MODELS THAT LEARN TO READ ALOUD

English orthography satisfies all the requirements of the canons of reputability under the law of conspicuous waste. It is archaic, cumbrous and ineffective; its acquisition consumes much time and effort; failure to acquire it is easy of detection.

—Veblen

[2]Since this writing, there have been several unpublished attempts to meet some of the prevailing criticism. Marchman and Plunkett (1989) argue that regression in performance of single units during training is a natural consequence of connectionist systems with multiple connectivity. A unit may initially be responsive to a particular input, only to lose that unique sensitivity as more trials increase the number of dimensions to be discriminated. Hence, individual units seem to "regress" simply as a function of increased training, without special changes in the model's input. There are several aspects of this work that mitigates its implications for natural developmental regressions. First, the loss of response discrimination found by Marchman and Plunkett is in *single* units, not a whole learning system; second, it is hard to show that the early success and later regressions in single-unit responses is more than statistical noise during early training phases.

MacWhinney and Leinbach (1991) have offered a broad range response to previous critiques of R&M. Their primary method is to construct a new model which acquires past tense morphology allegedly without TRICS. This model is presented somewhat cryptically, so analysing it is difficult. Its salient properties are (1) it is a four-layered model with two cascaded hidden layers; (2) it implements an "onset and rime" representation of sequentially ordered segments with a full complement of features describing each segment; (3) cases for training are presented with their actual relative frequencies; (4) some measure of success is claimed, but the model fails specifically to show a regression on irregular forms. In response, I note (1) two middle layers were used because it did not work with one layer—this is an interesting, possibly important, result bearing on the power of such models to learn sequence transformations. But without some analysis of the hidden units' performance the meaning of any learning is obscure—at best, it is an existence demonstration that some network configuration can learn aspects of orderly data; (2) They admit that the representations are those of standard linguistic analysis, but fail (explicitly) to see that these representations are *not* theory-neutral; rather, they have built into them reflections of how segmental and morphological structure work. Curiously they recognise this, claiming that theirs, and other, models are "implementations" of actual processes, and are therefore immune to independent criticism. This amazing point begs the question: If the models are merely implementations, then they make no claims and hence are of little theoretical interest, specifically connectionist. (3) It is intriguing that presentation of actual relative frequencies worked—indeed, surprising that it was possible in a tractable number of trials given that the frequency ratio between the most and least frequent verb is of the order of 10^4:1. It may turn out that the additional hidden layer is important in filtering out dominant effects of the overwhelmingly frequent irregular verbs. (4) They tout the model as achieving two out of three phases of U-shaped development, namely a phase of poor performance and a subsequent phase of good performance. A regression requires by definition a decrease in performance; accordingly, the most charitable interpretation of the boast that "two out of three isn't bad" is as a joke.

Foreigners always spell better than they pronounce.

—Twain

I want that glib and oily art
to speak and purpose not

—Shakespeare (King Lear)

The R&M model is relatively simple in that it uses a single layer of input nodes and a single layer of output nodes. Obviously, if this is supposed to be like a model of even a minuscule part of the brain, it is much too simple. Also, it is well known that a two-layer activation system cannot learn disjunctive mappings (Minsky & Pappert, 1969). One of the legacies of taxonomic linguistics as one level of representation in language involves disjunctive relations with units at other levels (as in "complementary distribution"). So it is unrealistic to attempt any kind of linguistic modelling with a two-layer system. More recent implementations of connectionist models include an intermediate level of nodes, so-called "hidden units", which do have the power to learn disjunctive mappings. In a scheme of this kind, there are two sets of connection strengths to adjust on each training trial; those between the output nodes and the hidden units, and those between the hidden units and the input nodes. Obviously, little would be gained by setting each layer of connections in the same way. There are a number of schemes for how to apportion the credit and blame to the two levels. A popular technique is so-called "backpropagation". In this scheme, the connections between the output nodes and hidden nodes are adjusted first on each trial, in a manner similar to that of a two-level system. Then the connections between the input nodes and the hidden units are adjusted (the way this is done is not in itself straightforward—it involves integrating the error term over all the connections from an input node to each of the hidden units. But it is a mathematically plausible extrapolation from what happens in two-layer systems. See Rumelhart et al., 1986).

Several models of different kinds of language learning have been developed using models with hidden units. Some of these models appear to learn to read, that is, they learn to pronounce English words, given conventionally spelled input. These models have received a lot of public attention and are worth some consideration as they are applied to linguistic problems.

One would like a spelling system in which letters always correspond to the same sound, but the relation between spelling and pronunciation in English is complex. First, there is a distinction between "regular" and "irregular" spelling correspondences. Even the regular correspondences are not simple, however: Most consonants actually have a small set of phones to which they correspond, and there are also consonantal doublets that map onto single sounds; vowels are pronounced in several ways, with a particular set of changes (from the Old

English vowel shift), being signalled by a final "e" in the next syllable. Part of the reason for this melange is historical—the alphabet was formed before a number of sound changes took place. It is also due to the fact that the spelling corresponds to the underlying morphological structure, which is itself computationally prior to the operation of phonological rules (see Chomsky & Halle, 1968).

Some enthusiastic researchers have referred to the English system as containing spelling-sound "rules" (see, e.g., Venezky, 1970), but nobody has been able to make a rule-based pronunciation program that really works—and the likes of IBM have thrown millions of dollars at the problem. But there are a number of majority regularities, and one can show that children follow them during early stages of reading, pronouncing relatively irregular words as though they were regular. There is also some evidence that brain-damaged adults revert to regular pronunciations of irregular words. A barrage of experimental studies with normal adults has shown that spelling irregularity increases reading time for infrequent words, but not for frequent words. Finally, subjects seem to have access to different strategies for deciding when a letter sequence is a word as a function of the other words in the experiment (see Seidenberg, 1987, for a representative review of these research areas).

All this (and more) has led an army of psychologists to agree that people read words with a "dual route" system (see, e.g., Coltheart, Sartori, & Job, 1986): Frequent words are memorised as a visual whole, whereas infrequent words are sounded out following the regularities (and then corrected if they are actually irregular). With a relatively regular and frequent set of words to decide on, subjects can decide directly on the visual appearance; if the set of words contains rare irregulars, then the subject has to sound out each word to check its legitimacy as a word.

Psychologists persist in believing that the mind is simple: They ponder a two-process model like that just described, and wonder, "why not *one* process?" In this vein, Seidenberg and McClelland (1988; S&M) have constructed a connectionist model which learns to map English monosyllabic uninflected words onto appropriate Wickelphonological feature matrices. The model has three layers of nodes (see Fig. 8.2). The input level is a set of 400 nodes, which represent an encoding of letters. Each input node is connected to 200 hidden units. The output level has two sets of nodes; one is the set of 460 Wickelfeatures as used in the past-tense verb learning model. The other is a set of nodes which encode the letters the same way as in the input nodes. The model was trained for 250 cycles that selected 400–500 words from the 2900 most frequent monosyllabic English words. Words were selected on each cycle in proportion to the log of their real frequency of occurrence (this was accomplished by a random process on each cycle, which is why the number of words on each cycle could differ slightly).

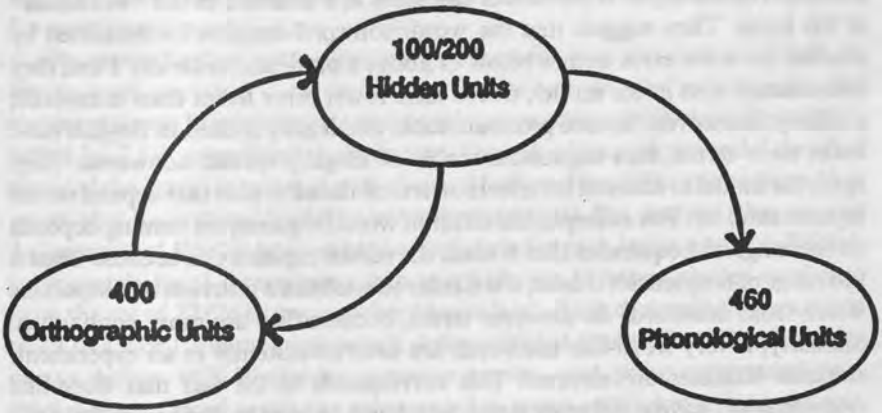


FIG. 8.2. Overall structure of the learning model used by Seidenberg and McClelland (1989).

On each trial, the feedback information was both on the correctness of the Wickelfeatures that were and were not activated, and on the correctness of the encoded letter representations. S&M measure the success of the model in terms of the sum of the squares of the difference between the activation level of each output node and the level it ought to have (0 or 1). The model shows considerable improvement in these terms over the 150,000 individual word trials. Quantitatively, the model learns the frequent regulars and irregulars first, and learns the infrequent regulars last. S&M note that this is similar to the order of learning in children. They also quantify the accuracy of the model after the 150,000 trials and find that the mean squared error is not larger for irregular than regular words unless they are also infrequent: They relate this to the experimental findings on word naming and lexical decision. They show that the model's error score on certain subsets of words corresponds to their relative time to be read aloud in naming experiments. They simulate phonological priming by giving an already trained model *additional* training on a word, and showing that training lowers the error score for other words with similar spelling-sound correspondences. Finally, they train a new model with few hidden units and show that it only learns to pronounce regulars and frequent irregulars, just like developmental dyslexics.

S&M also describe an interpretation of the model that accounts for effects in "lexical decision" studies, a paradigm in which subjects must decide whether a letter sequence is a word or not. S&M suggest that the error term on the re-

activation of the input Wickelnodes can serve as a measure of the "wordiness" of the input. They suggest that the word/nonword decision be simulated by whether the letter error term is below or above a particular criterion. First, they demonstrate that in the model, words have lower error terms than nonwords; similarly, nonwords that are pronounceable and legally spelled in English have lower error terms than unpronounceable or illegally spelled nonwords. They apply the model to account for effects on lexical decision time that depend on the experimental set. For example, the effect of word frequency on naming depends on the range of frequencies that is used: the model explains this because when a full range of frequencies is used, it is harder to establish a criterion that separates words from nonwords on the error terms, because they are more continuous. Similarly, if very word-like nonwords are used consistently in an experiment, response latencies are elevated: This corresponds to the fact that word-like nonwords will have small error terms, which will be hard to distinguish from the error terms for real words.

This model is a tour de force. The creators are modest about many aspects of the model, and enjoin us not to take specifics of it too seriously. But they emphasise that the model appears to do away with *two* routes for reading and to replace them with *one*. Seidenberg (1989) also suggests that the model exemplifies how a continuous framework can represent what appears to be categorical knowledge.

But this claim may be premature—just as in R&M, the model works because of the real regularities of the input/output correspondences and because of a set of special devices not dissimilar to those in R&M. First, since it uses the Wickelphonological nodes from R&M, it also buys all the same representational moves (except blurring, which they did not use—there was no need to have a phonology-hearing representation, since they were not attempting to account for phonologically based misreadings). These devices conspire to emphasise the central phone information, which facilitates the learning of individual letter-phone correspondences in regular words. They also emphasise the clarity of initial and final phones, which help differentiate the large number of letters which have a different distribution of pronunciation in word initial and word final position (e.g. y,h,w,c,e,i,q,u,x). Finally, in monosyllabic words, which the model was limited to, consonantal letter combinations almost all occur at word boundaries: Hence giving extra marking to those sounds which break the one-letter/one-sound generalisation.

The use of Wickelphonology also makes predictions about the different ease of learning to pronounce words as a function of the number of distinct Wickelfeatures in their description. For example, since "dozed" has 55% more Wickelfeatures for the medial sounds than "fazed", there should be an enormous difference in the model's accuracy on those words. Given S&M's method of measuring success, "fazed" should be learned sooner and better, but that difference does not correspond to any real effects that have been reported

(actually, S&M do not work with inflected forms, but similar points can be made for uninflected words.)

The second set of representational devices involves the way in which the letters were encoded onto input nodes. Because of the requirement that representations be distributed, they used an encoding of Wickel-letter triples, with L1,L2,L3, as outlined in (6). It might have broken new ground if they had encoded the letters in terms of a set of visual features that differentiate them (they report that Lacouture [1988] has recently done that). But, instead, they created 3-dimensional Wickel-letter-matrices as labels for each input node. A Wickel-letter-matrix has 3 dimensions, each of which has 10 letters chosen randomly from the set of 27 (26 letters and word boundary). Each dimension corresponds to L1 or L2 or L3 in the triple which defines Wickel-letters. Each Wickel-letter-matrix defines 1000 Wickel-letter-matrix points—each point corresponds to a Wickel-letter. Input words are represented in terms of Wickel-letters—a given input node is triggered if any of the 1000 Wickel-letter-matrix-points correspond to any of the Wickel-letters in the input word. Four hundred input Wickel-letter-matrices were created by randomly selecting 10 letters for each of the 3 dimensions 400 times.

This encoding of the letter appears at first to be computationally neutral, except there is a twist: The matrix dimension corresponding to the middle letter of Wickel-letters never has word boundary chosen as one of its ten letters. Prima facie, this is a sensible property, since, by definition, no Wickel-letter of the form X#Y exists in monosyllabic words. But it also has the property that the central letter of Wickel-letters is represented in ~4% more input node Wickel-letter-matrices than the initial or final letter of Wickel-letters. This is a small difference, but the model does run for 50,000 individual trials, so it is likely that it enhances the relevance of the central member of the Wickel-letters, which is what would be needed to capture those regular letter/sound relations which exist. Also, there is an organisational feature of the feedback that may be critical to its success. Each hidden unit is trained simultaneously with feedback from the letter and phonological nodes. Insofar as there are spelling/sound regularities, this technique allows the hidden units to respond in accord with correlations between letter-sound pairs: Each unit is given feedback on each Wickel-letter-matrix-node Wickelfeature pair, which will tend to isolate the training effect of those pairs which correlate (be it positively or negatively).

As in R&M, the training words are not presented in direct proportion to the actual frequency of word occurrence. In English, about 180 words account for more than half of the word occurrences: The range across all words is at least 70,000 to one. Among monosyllabic words, the 180 most frequent words account for more than 90% of all word occurrences, because most infrequent words are also multisyllabic: More than 97% of the 180 most frequent words are monosyllabic, whereas less than 20% of the least frequent words are monosyllabic (this can be determined from analysis of the Kuçera & Francis [1967] word

count). Hence, if words were presented in linear proportion to their real frequency of occurrence, the model might well learn the more frequent irregular words as isolated cases, and never extract regular patterns or differentiate them from the irregulars—a result not consistent with the facts S&M want to simulate. Furthermore, it appears to be the case that irregular spellings are more probable in frequent than infrequent monosyllabic words. Thirty percent of the 90 most frequent monosyllabic words are irregularly spelled, whereas this is true of only 7% of a corresponding sample of the least probable monosyllabic words. Hence, irregular pronunciations would swamp the learning system if words were presented with their actual relative frequencies.

For these reasons, reducing the contrast in exposure to the most and least frequent words was a practical necessity. As I mentioned in the discussion of R&M, this manipulation removes the model from the realm of plausible learning models; at the same time, it converges on what a structural analyst would do when searching for possible spelling/sound patterns. In this case, "McSeidenwell's Demon" collapses the frequency range of tokens via a log transform. Clearly, it had to leave some frequency information intact (unlike McRumelwell's Demon, who manipulated the R&M model by changing its input on cycle 11), because it is frequency-based human behaviour which they are attempting to model. A log transform may have been chosen because it flattens the distribution enough, but also leaves some frequency information available. One might argue that a log transform of actual frequency is justified because the effects of experience often turn out to be a log transform of frequency of training. But, if true, that fact is to be *explained* by the learning model, not artificially built in to the data presentation scheme by a word-knowing demon.

In response to an early version of this paper, Seidenberg and McClelland (1989) tried out a variant of their model in which words are presented more closely in proportion to their real frequencies. Instead of compressing the frequency range with a log transform, they tried a square-root transform. With training, this model displayed the effect of frequency, which S&M note with satisfaction. They also note that the model does *not* display the facilitation of high-frequency regular words, although they do not explain this fact. Furthermore, they do not note that the low-frequency regular words are learned relatively less well. Table 8.2 presents the mean squared error for each of the word categories for about the first 25,000 training trials (the first 40 epochs for the original model, taken from their Fig. 3, and the first 400 epochs for the square-root transform model, taken from their Fig. 27).

The most noticeable fact is that low-frequency regular words are learned relatively less well in the root than the log model. This is *exactly* the kind of difference predicted by the preceding consideration: The high-frequency irregulars occur relatively more often, and drive out the regulars. The root model maintains considerable reduction of the real numerical superiority of the high-frequency irregulars. It remains the case that learning all regulars will be less

TABLE 8.2
Approximate Mean Square Error from Seidenberg and McClelland (1989) Comparing Models in which Words are Presented with a Frequency that is Either the Log or Square Root of their Actual Frequency

	High Frequency		*Low Frequency*	
	Regular	*Irregular*	*Regular*	*Irregular*
Log Model	7	10	10	15
Root Model	7	8	12	16

effective if real frequencies are used. Of course, with enough training, or enough hidden units, some model will learn all words to some degree. The important fact is that the relative ease of learning will be shifted in favour of irregulars if real frequencies are used. This is disastrous for S&M, because the relative ease of learning different word types is the basis for making predictions about behavioural complexity.

S&M give a scientifically oblique argument for using the log frequency compression: Without it, the model would have to be given 5,000,000 training trials to guarantee training on each word—the relatively frequent words would take up most of the training trials. They estimate that 5,000,000 trials would take about a month with existing equipment. Given the amount of human brain time they and others are spending on their model, the month might have been well spent: Most of the behavioural experiments they model took at least that long to run.

They also offer several potentially empirical justifications for some kind of frequency compression: First, they suggest that infrequent regular words appear more often in inflected variants (less characteristic of some irregularly spelled words such as "have"); hence their true relative frequency is underestimated. Second, they speculate that the frequency range is smaller for children learning to read. The first point could be determined empirically, but since most irregular words are also constant when inflected, it seems unlikely that the frequency levelling effect of inflected forms is very large. The second point is more puzzling. It is true *by definition* that children experience a smaller frequency range: Until the child has read at least 70,001 words, it has no potential to experience the true frequency range of English words. But the child also does not experience most of the individual words at an early stage. So the relevance of the child for the model is obscure. Of course, it would be easy to use a children's text-based frequency count for selecting training trials. I made a small test count of 2,000 sentences of a second-grade text. Irregularly spelled words accounted for more than half of the forms. Furthermore, the mean token frequency of the irregulars was about 30 and of the regulars was 3. This is necessarily a smaller frequency

ratio than for adults at least because of the small samples, but it suggests a similar problem.

In brief, S&M's model does just what its designers created and manipulated it to do—it learns spelling/sound regularities, and it learns frequent irregulars more quickly than infrequent irregulars. But, of course, S&M aspire to show that once trained, the model makes distinctions along the frequency and regularity dimensions similar to adult humans: They use the mean squared error between the primed features and the actually correct features as a way of predicting relative reaction times. This presents a problem if one is supposed to take the model as a candidate for reading behaviour: The mean squared error of an output can be calculated only if the correct pronunciation is already known. There is a corresponding circularity in the presumed behavioural model: The reader transforms the letters into the phone-features, and then checks it against his knowledge of the correct pronunciation of the word—the further away the actual word is, the longer the time to pronounce the word. But if the reader "knows" the correct pronunciation of the word, why doesn't he just say it? There might be a way out of this: One might try to assess the distance between the output of the model on a word and the nearest pronounceable sequence (assuming some English phonology filter that defines "pronounceability"—a neat trick in connectionist terms). It *might* turn out to be the case that those words which arrive more closely to a pronounceable sequence are the short and frequent words . . . but that would have to be shown. In any event, S&M make much of the fact there is no word-level set of units, so it is totally mysterious how an error term could be matched against anything at all within the model.

The lack of a transparent relation between the model and a behavioural mechanism further highlights the fact that the learning net is telling us only about redundancies and associative patterns available in the input (once treated to some statistical distortion). That is, it serves as a powerful frequency analyser, operating on a predigested set of actual data. Perhaps this may reveal useful statistical properties of English spelling/sound correspondences; but it is not a model of pronunciation behaviour. Hence, among other things, the claim that it is a "one-process" model of pronunciation is unwarranted.

The "simulation" of lexical decision does not involve a behavioural implementation, because it does not aspire to be a simulation of actual behaviour. Rather, the re-activation letter error term is used as a measure of the definiteness with which a word is recognised. All the predictions are based on the clarity with which different subsets of words and nonwords can be distinguished using the letter error term. Thus, the range of phenomena that the model accounts for are just those which have to do with the statistical properties governing the discriminability of different subsets of English words. I have no quarrel with this application of the model: Indeed, it demonstrates elegantly that subjects use the statistical properties of letter sequences to guide their lexical decisions. It makes no claims about how that process occurs.

The model's statistical success might be taken as a demonstration that lexical decision may not be a very useful task for the study of language behaviour—if the model were a complete account of lexical decision behaviour, using it to study semantic structures or sentence processing would be like studying taste by using the amount of time it takes people to name pictures of vegetables. Also, the flavour of S&M's discussion is redolent of the view that semantic factors are ordinarily unused in lexical decision: Many studies, however, show that associative and other kinds of priming affect lexical decisions (classically, Meyer & Schvaneveldt, 1971, and many others since then). In fact, lexical decision may be sensitive even to structurally mediated semantic information: For example, lexical decisions are faster following a noun-anaphor that refers to the word (Cloitre & Bever, 1986). S&M do not directly deny such possibilities, but the model is irrelevant to them. And, since the model has no word-level representations, it is difficult to see how to integrate it with semantic and syntactic effects in the normal uses of language.

The most striking empirical success of the model is the orderly relation between the phonological error score and actual reaction times in lexical decision tasks: The model predicts with a high correlation the relative decision times for 14 different subsets of words (e.g. "frequent regulars," "infrequent regulars," "infrequent uniquely spelled words"). This correlational success is an initial demonstration that subjects are responsive to statistical regularities that differentiate spelling-sound correspondences in different types of words. But it leaves unspecified how subjects' average performance is accumulated from individual responses. One appeal of the connectionist models is based on the appearance that they reduce apparently complex phenomena to a single process—spreading activation among units. Accordingly, this model's statistical success appears to confirm the hypothesis that subjects also use a single activation process in lexical decision. This conclusion, however, is unwarranted, if for no other reason than that the model is not a behavioural theory.

The ambiguity of the data here parallels that of studies of "one-trial vs. incremental learning" and "probability matching vs. hypothesis formation". In each case, the data overall have an incremental appearance, gradually increasing as a function of number of trials: But closer analysis suggests that the averaging process is obscuring quantal learning acquired by different subjects at different times. There is a directly relevant experimental analogue to consider in the present case—the acquisition of the pronunciation of new written words by young children. Suppose one attempted to teach a young child to read vocabulary in a new language: On each trial, the child is presented with a word, attempts to pronounce it and is given the correct pronunciation as feedback. With a vocabulary of 100 words or so, 75 with a "regular" pronunciation and 25 with a set of irregularities, one would expect to see a gradual improvement across words and children, and such improvement should occur faster with regular pronunciations (assuming equal presentation frequency). But it would not be

surprising if individual children learn the pronunciation of individual words quantally, that is, going from not pronouncing a word correctly on trial n to pronouncing it correctly on trial $n+1$, and never misreading it after that. Across children, n should turn out to be greater for irregulars, because on average, their pronunciation is less supported; for any individual word, the acquisition will appear gradual, because the function is averaged across individual children.

In this way, a process that is actually discrete can appear to be continuous: A difference between word classes that is quantal can appear to be gradual. The corresponding possibility exists for tasks such as lexical decision in adults. Each trial may be responded to quantally, with a discrete process: For example, "assume the word is frequent and can be checked automatically; if it isn't, assume that it is spelled regularly. . . ." Across words of different kinds, the average response times will correspond to the extent to which the particular kind of word follows the more favoured strategies. Hence, a multi-strategy model fits the data exactly as well as an alleged "one-process" model. In the absence of a particular behavioural theory, the data and the success of the model do not bear on the behaviour, only on the statistical regularities with which the real behavioural mechanisms are interacting.

There is another, more infamous, model which learns to pronounce English words, also using hidden units. This model, NEttalk, is trained to transform printed words into input phone instructions to a Dectalk machine (Sejnowski & Rosenberg, 1986). The model is trained on each letter, moving from left to right; on each trial, the training occurs on a letter with the 3 preceding and 3 following letters. One can think of this as training on super Wickel-letters, with 3 pre- and 3 post-central phone positions. (In this case each Superwickel-letter is not given its own node, since that would add up to the equivalent of 27 to the 7th power Superwickel-letter nodes. In fact, the relative order of the 6 surrounding letters is not preserved in the input representation.) After 400,000–1,000,000 trials, the model-driven Dectalker sounds very impressive.

How does this work? First, the model is learning the associative information available, which determines the sound of a central phone in a very informative string (7 letters). Second, it dodges the problem of digraphs, by requiring the model to output silent phones for one of the members of the digraph (e.g. "phone" would have at its output level "f-on-".) This solves artificially a big problem in spelling/sound correspondences. Finally, Dectalk itself has a program with a great deal of English phonology built in. For example, it fixes up phonetic transitions, has a lot of stress rules, and includes many phonetic rules of English.

It is striking that both of these models of "reading" miss the real computational problem in reading. Written text provides an input sketch of what to say, but the output is determined not only by the letter sequence but by computational linguistic processes as well. For example, it is well known that the morphological and surface phrase structure of words plays a large role in stress,

and therefore in such phonetic features as vowel length, and neutralisation of both vowels and consonants: S&M avoid this problem (a) by not having their model responsible for it and (b) only treating monosyllabic words. Netalk solves it (insofar as it does) by relying on long string of input information and on the rich built-in phonology in Dectalk. But it makes the obvious mistakes which a nonmorphologically aware reader makes, such as not differentiating the pronunciation of "ragged" in the two sequences in (11), or of "delegates" in (12).

11. a. Harry's ragged.
    b. Harry ragged.
12. a. Harry delegates minors in gym.
    b. Harry's delegates minor in gym.

In brief, insofar as these reading models work, they do so because they limit themselves to that information which is indeed associatively determinable, and rely on other devices to pick up some of the pieces.

These models require perfect information and many thousands of trials, even with all the special devices and tricks. Clearly, they are able to extract certain frequency-based information in the input/output relations they are trained on. This may be marvellous engineering; it may lead to conversations with telephone operator modules as intelligent as the current real ones. It may reveal to us frequency properties in the world of which we were unaware. But the value of psychology is limited to knowing that a powerful device shows some adaptation to some frequency information after 100,000 trials involving 400,000 instances of phones and billions of individually computed adjustments in associative strength (e.g. in S&M's model there are 172,000 connections, each of which must be checked and may be changed on each of 10,000 trials. So the training period involves more than 200 billion computations.). Science has given us many analytic tools that tell us about the world, but not directly about ourselves. A spectrometer tells us about the composition of a distant star—but that does not tell us anything about stargazing. Aside from all the limitations and isolated cleverness of current connectionism, maybe what it is showing us is that this cannot be the way that children learn anything, even habits.

One might think that the very large number of trials involved in connectionist models is irrelevant because the human child has so many more units to learn with: Surely, if a model with 200 hidden units requires 100,000 training trials, one with 2,000 will require fewer trials, with 2,000,000 fewer still, and so on. This is undoubtedly true. But the computational load (the number of individual computations and weight changes) on each trial also increases proportionally. Most telling, however, is the fact that as the number of hidden units goes up, the models tend to learn each case as a separate item. Hence, the models converge onto brute-force one-to-one mapping associations that cannot generalise to new cases. So, the prognosis is not a hopeful one: When you make the models

powerful enough to deal with more than toys, they memorise everything as a separate instance.

THE STUDY OF STATISTICALLY AVAILABLE INFORMATION

How use doth breed a habit in a man!
—Shakespeare (Two gentlemenof Verona)

Most people don't associate anything—their ideas just roll about like so many dry peas . . . but once you begin lettin' 'em string their peas into a necklace, it's goin' to be strong enough to hang you, what?
Lord Peter Wimsey

Structural Hypotheses and their Statistical Confirmation

I originally observed that connectionist modelling might help us understand the relation between computational and habitual mental structures. This is justified by what I found, but it is difficult to show that from most existing models. This is because these models make a theoretical virtue of obscuring the structural information they contain and presuppose. Indeed, the flavour of the accompanying commercials is that no structural information is required at all. That makes it a chore to show how the models really work, and it reduces the chances that they will tell us anything about the real relations between the structural demon and the statistical beast. This is consistent with the fact that some of the models' creators seem to argue (like Bates & MacWhinney, 1987) that the beast *is* the demon—that is, that the notion of structural rule and category is actually an illusion, and that conceptual nodes and associative strengths between nodes are all there is.

Proponents of this kind of view (e.g. Rumelhart and McClelland) accept the scientific utility of rule-based descriptions of language; but, they argue, rules express vague regularities inherent to the data, whereas connections express what the underlying mechanism actually is. A frequent explanatory allusion is to the relation between Boyle's law (BL) and the internal combustion engine. The idea seems to be that BL is a generalisation about the behaviour of pressure, volume, and temperature, but the engine is a real instance of those generalities. This analogue has been offered to me a number of times, so it may be worth a little direct attention. The analogy is clearly flawed, but perhaps in an instructive way.

The main flaw is that it begs the question about ultimate cause while seeming to answer it. BL states an orderly relation between pressure, volume, and temperature. It exists for a set of reasons—a combination of thermodynamic

laws and assumptions about the particulate nature of gases. That is, BL is true in an engine because of a set of laws not contained in the engine. If structural linguistic rules are true of language, then either they are in the language engine or not. If they are not, then they must be true for other reasons, presumably materialistic ones. Hence, just as for BL, there must be other physical principles that explain the existence of the rules, not represented in the connectionist model itself. This would argue that there is some other biological basis for the rules, not included in the connectionist model, but which constrains it to behave as though it included them. So the question is begged.

The currency of connectionist learning is statistical inference. We know, however, that no amount of inference can cause the distinction between something that is always true in the model's experience, and something that is true by categorical designation. Examples of this kind of distinction are not hard to find. National Basketball Association players are tall in *fact*, but they are professional in *principle*. Such categorical distinctions are not mere artefacts of a technological civilisation (assuming professional sport to be civilised): In primitive cultures, it can be a matter of great importance to know who your biological ancestors are and who merely acts like them—it determines everything ranging from politeness forms to marriage. Even in civilised culture, the mere knowledge of such categorical relationships can be a matter of life and death—after all, Oedipus behaved perfectly appropriately with his wife: The basis of the tragedy was that he insisted on *finding out* that she was his mother, very much the wrong category for a wife.

There are varying responses from connectionists concerning the problem of explaining categorical and symbolic knowledge: The most apparently forthcoming is that such knowledge could be built into the models—which is why they are allegedly consistent with nativism. But, insofar as categorical distinctions and symbolic constructs are built in, the connectionist model becomes merely an implementational system of constraints to be explained by some other theory. One such "explanation" for innate categories that the more enlightened connectionists offer is "evolution." This surely must be the true account of whatever is innate. But appeals to "evolution" do not explain the mental nature or role of categorical and symbolic processes, much less why the particular ones have evolved the way they have. It is another example of begging the question. On one view, the constraints leading to the possibility and formation of categories are biological, in which case they are not explained by connectionist models. On the other view, the constraints are metaphysical (e.g. constraints or information processing systems in general). The final possibility is one that I suspect many connectionists will find most satisfactory: There is no explanation for the existence or arrangement of categorical mental structures; there is only biological history.

There is a corresponding problem in accounting for the acquisition of complex systems like language via statistical inference alone: Inference is a

mechanism for the confirmation of an hypothesis, but it is a notoriously bad mechanism for *creating* hypotheses about the kinds of structures inherent to human behaviours. Consider a (relatively) simple phenomenon—learning to play tag—in particular, figuring out the concept of "it." (See Bever, Carroll, & Miller, 1984; Lachter & Bever, 1988.) "It" is actually a category defined by one of the recursive rules of the game, not by any particular overt property or necessary behaviour. It seems likely that children can learn this concept without much explicit instruction. But the statistics of the behaviour of the players is not informative enough to account for the creation of the categories: Children have to have a lot of conceptual information, which sets up hypotheses as to what kind of activity they are watching. They have to know about games; they have to know about transitivity of (invisible) properties (like "it"-ness); they have to know about rules and priorities among them. Such prior knowledge allows for the construction of hypotheses about tag. Once the alternative hypotheses are elaborated, the statistical properties exhibited in games of tag provide information that can tend to confirm or disconfirm particular hypotheses.

It would be unreasonable to assume that the child uses none of the statistically available information to confirm its hypotheses. I have argued that some connectionist models of learning work only insofar as their structure (and input data) steer their associative processes towards correct representations. Thus, a successful connectionist model is actually a hybrid between associative and symbolic structures. This dual property may make it possible to investigate the interaction between what is built into a learning model, and the environmental regularities which confirm its hypothesis. In this way, we can separate the symbolic process of forming structural hypotheses from the statistical processes involved in two aspects of language acquisition. First, the statistically supported patterns can confirm specific structural hypotheses. Second, they can become the behavioural bases for extending linguistic performance beyond the current computational capacity. In the next pages, I outline an example of a connectionist model of learning phrase segmentation, and show how the knowledge it isolates might serve each function.

The problem of language acquisition is often framed as the problem of learning the grammatical structures that generate a set of experienced utterances stored with their meaning. On this view, children accumulate a set of sentences they understand, and then apply analytic processes to them which yield a grammatical organisation for them. This view enables theorists to focus on the kind of learning model required for the discovery of a grammar from coded input data, and has been a fruitful strategy. But it also begs an important question: How do children assign both linguistic structure and meaning to sentences for which they have not yet learned a grammar? (See Valian, 1990, for a discussion of this issue and its impact on parameter setting theories of language acquisition). That is, how do children encode sentences they cannot parse? A related question is: How do children make use of partially correct analyses of the utterances they hear?

Consider the role of connectionist modelling in the context of an hypothesis testing model of language acquisition. On this model, children have competing structural hypotheses about their language and use incoming data to choose between the hypotheses. Children could use such data in two ways: Deductively, the possibility of a particular utterance can rule out a structural hypothesis; inductively, accumulation of partial information can confirm one hypothesis statistically over another. Deductive models of hypothesis testing in language acquisition have been the most commonly explored type—children are presented as "little linguists," applying information about which sequences are grammatical to distinguish underlying grammatical hypotheses. A great virtue of such models is that they can be studied in direct relation to a grammar that distinguishes between grammatical and ungrammatical sequences—thus, the learning model is tied directly to the grammar. However, deductive models generally require a body of structurally analysed and comprehended data, and hence beg the question of how sentences that cannot be parsed are understood and recorded as data.

Inductive models have appeared unattractive for three reasons. First, it is axiomatic that pure inductive systems are unable to form hypotheses, so a separate hypothesis-formation process must be postulated anyway. Second, it is not clear how partially coded input could bear on structural hypotheses. Third, it is difficult to assess the availability to children of statistically supported hypotheses. If we interpret connectionist models as inductive analysers, then we can use them to meet the second and third objections to inductive models of hypothesis confirmation. Consider the following general description of connectionist learning and what it can show:

> A model has input data, *I*, under a particular descriptive scheme, *R*; and output data *O*, presented in a set of behaviours, *B*. The model is trained to produce a discriminative response *DB* that discriminates along some dimension of *R*, *DR*. Insofar as the model learns to pair instances of *DB* and *DR* correctly on a training set, it can be asserted that *R* provides the information needed for the discrimination in the training set; insofar as the trained model correctly discriminates new *I*, it can be asserted that *R* provides the information needed for the *DB*/*DR* discrimination in general.

Or, to put it succinctly, a model can reveal the *cue validity* of information for a discriminative behaviour, under specific descriptions of the environmental input and the behavioural output. On this view, such models are analytic tools for the psychologist, aiding in the discovery of what information lies in the statistical properties of the environment under different descriptions.

The fact that models can operate under different input descriptions allows us to construct an input representation for sentences that is incomplete, and then examine the cue validity of that kind of representation for certain kinds of information children receive. Consider the following model, "Baby Clauseau,"

which learns to predict when an utterance is going to end (see Juliano & Bever, 1990, for a fuller presentation). It has the following assumptions:

1. Input: examines three words of input at a time. Each word position has:
   a. a separate node for each of the 100 most frequent words in motherese;
   b. a separate node for each of 4 word-length categories;
   c. a node indicating how far to the left the nearest recognised word is.
2. Output: an activation level of one node ranging from 0 to 1, corresponding to the likelihood that the third word ends the utterance.
3. Training feedback: negative feedback when the utterance does not end there; positive feedback when the utterance ends there.

We constructed a three-layer model with these assumptions, using 10 hidden units in the usual arrangement (complete connectivity between layers, a standard backpropagation weight-changing routine, etc.). It was trained on a 21,000 word sample of motherese text (approximately 4000 utterances), and then generalised to a new 800 word text of motherese.[3] The results were quite intriguing. First, the mean output level on the generalisation text for actual utterance boundaries was 0.5, whereas for non-boundary positions it was 0.1. This demonstrates a considerable degree of differentiation of actual boundaries, based on the minimal input assumptions. Second, if we separate the continuous output values into discrete predictions of boundary vs. non-boundary, the results appear as in Table 8.3.

Finally, we assessed the cumulative likelihood at each of 20 output activation values, of an actual utterance boundary (a "hit," a correct prediction by the model that there is an utterance boundary), and the cumulative likelihood at those values of a non-boundary (a "false alarm," or incorrect prediction of a boundary). We used an adaptation of detection theory to scale the discriminability of boundaries from non-boundaries as analysed by the model (Fig. 8.3).

TABLE 8.3
Percentage of Utterance Boundaries in Next Text by Baby Clauseau (BC), Using an Arbitrary Threshold Between .1 and .5 in BC's Output

	Model's Predicted Boundary
Actual Text	
Boundary	76
Non-boundary	18

[3]In fact, this model is a variant on one we constructed for a practical reason—to make on-line decisions about formatting text displayed in closed captioning for the deaf.

This technique is useful in assessing discriminability functions of such models, because it utilises information from every level of output. Figure 8.3 shows that there is considerable information at all levels of output that discriminates boundaries from non-boundaries, given the input.

This model demonstrates that a child could learn to predict when an utterance ends, knowing only 100 frequent words, the approximate length of all words and by examining three words at a time. Such knowledge can help the child delimit sentences in multi-sentence utterances without clear acoustic boundaries, clearly an important prerequisite to analysing sentence structure. But such multi-sentence utterances may be relatively rare, and hence the model might be taken a a small tour-de-force but not as an interesting one.

It becomes more interesting if we consider why the child might learn to predict when an utterance ends. The child and adult goal in comprehension includes

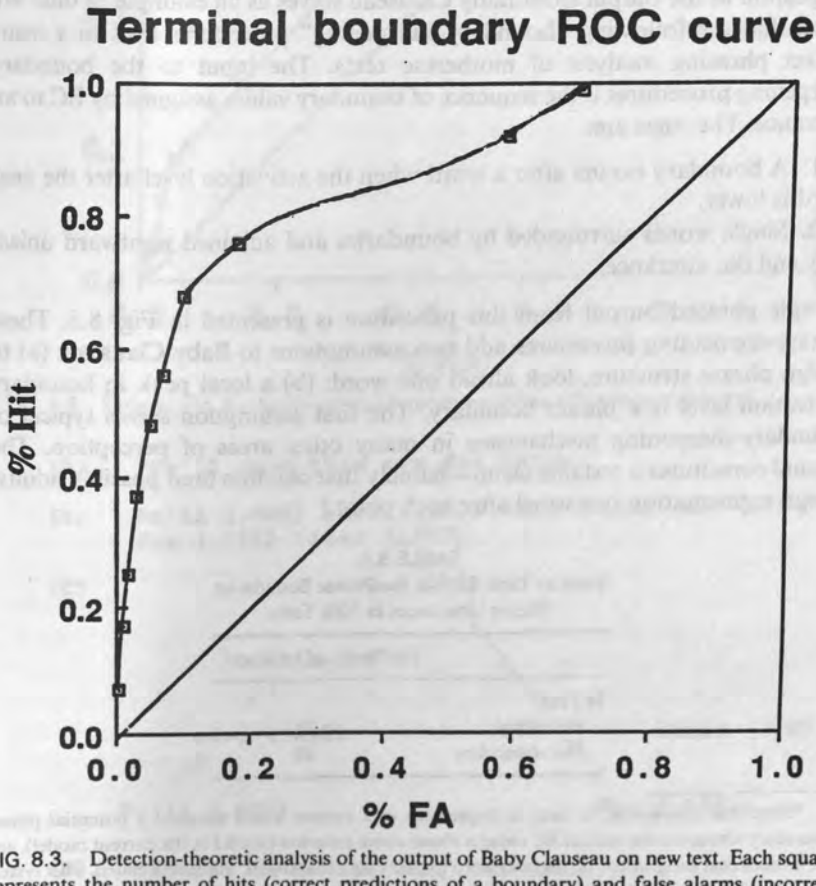


FIG. 8.3. Detection-theoretic analysis of the output of Baby Clauseau on new text. Each square represents the number of hits (correct predictions of a boundary) and false alarms (incorrect predictions) at successively higher output values.

segregation of major meaning-bearing phrases. Thus, the learning of utterance-final boundaries may be viewed as a by-product of the attempt to package sentences into meaning units in general. We can assess the significance of this interpretation by generalising the output of the trained model to predict phrase boundaries *within* utterances. We did this on the 800-word generalisation text period. The mean output value was 0.3 for within-utterance phrase boundaries and 0.1 for non-boundaries. Applying the same procedure as earlier to arrive at discrete predictions, we found the results as in Table 8.4.

Finally, the detection-theoretic analysis of the output shows considerable discrimination of phrase boundary from non-phrase boundary positions (Fig. 8.4).

I have argued that a separate function of associative knowledge is to extend the empirical domain of a computationally limited processing system. An adaptation of the output from Baby Clauseau serves as an example of this. We noted that the following "boundary sharpening" procedures lead to a near-perfect phrasing analysis of motherese texts. The input to the boundary sharpening procedures is the sequence of boundary values assigned by BC to an utterance. The rules are:

1. A boundary occurs after a word when the activation level after the next word is lower.
2. Single words surrounded by boundaries and adjoined rightward unless they end the utterance.

Sample phrased output from this procedure is presented in Fig. 8.5. These phrase-segmenting procedures add two assumptions to Baby Clauseau: (a) to assign phrase structure, look ahead one word; (b) a local peak in boundary activation level is a phrase boundary. The first assumption seems typical of boundary-sharpening mechanisms in many other areas of perception. The second constitutes a testable claim—namely that children (and possibly adults) assign segmentation one word after each point.[4]

TABLE 8.4
Same as Table 8.3, but for Phrase Boundaries Within Utterances in New Texts

	Predicted by Model
In Text	
Boundary	71
Non-boundary	20

[4]Note that this would be easy to implement in a system which assigned a potential phrase boundary whenever the critical BC value is above some criterion (say 0.2 in the current model), and then check that assignment on the next word to make sure that the BC value descended. This system would make empirically testable predictions about local garden paths, predictions which we are now starting to test on adults.

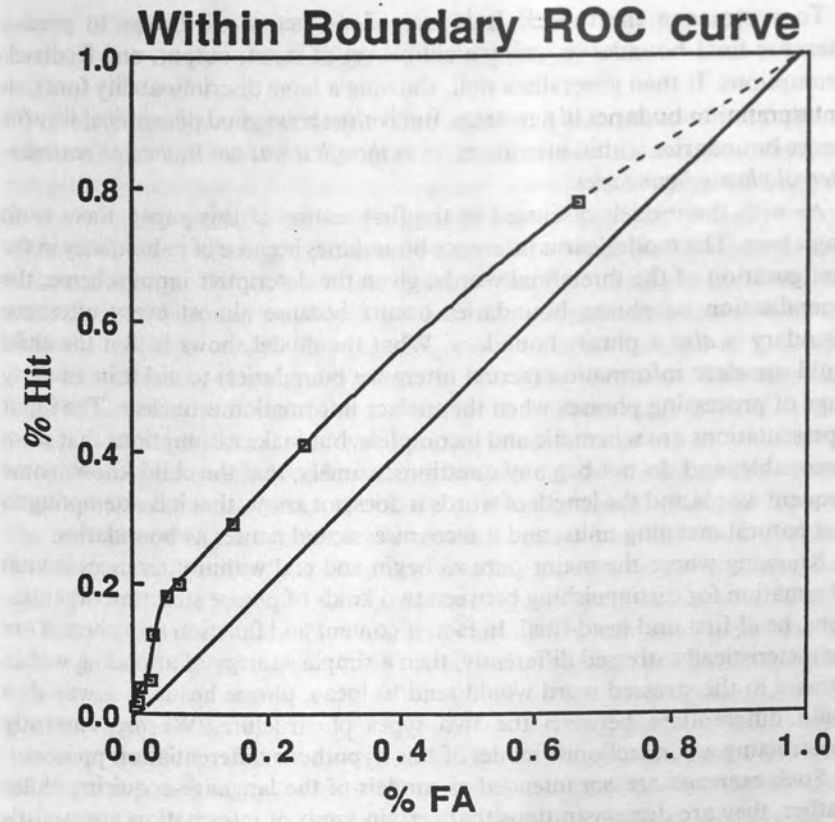
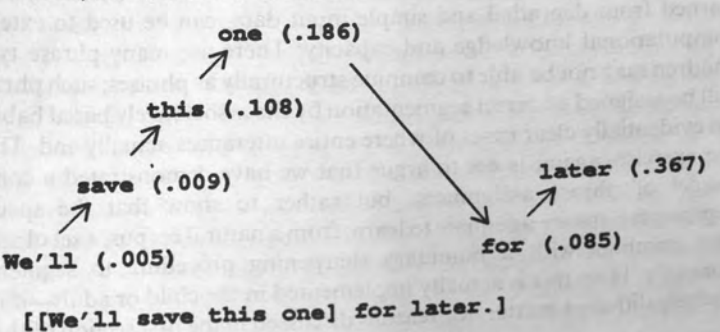


FIG. 8.4 Same as Fig. 8.3, but for phrase boundaries within utterances in new text.

(A) "We'll save this one for later."

(B) We'll (.005) save (.009) this (.108) one (.186) for (.085) later (.367).

(C)



(D) [[We'll save this one] for later.]

FIG. 8.5 Phrase segmentation using output from Baby Clauseau.

To summarise the model. It is trained on actual motherese to predict utterance-final boundaries, using a simple set of input, output, and feedback assumptions. It then generalises well, showing a large discriminability function for utterance boundaries in new texts. But it *also* shows good discriminability for phrase boundaries within utterances, *even though it was not trained on sentence-interval phrase boundaries.*

As with the models discussed in the first section of this paper, there is no magic here. The model learns utterance boundaries because of redundancy in the configuration of the three final words, given the descriptive input scheme; the generalisation to phrase boundaries occurs because almost every utterance boundary is *also* a phrase boundary. What the model shows is that the child could use clear information (actual utterance boundaries) to aid it in an early stage of processing phrases when the surface information is unclear. The input representations are schematic and incomplete, but make assumptions that seem reasonable, and do not beg any questions; namely, that the child knows some frequent words and the length of words it does not know, that it is attempting to find natural meaning units, and it recognises actual pauses as boundaries.

Knowing where the major phrases begin and end within utterances is vital information for distinguishing between two kinds of phrase structure organisations, head-first and head-final. In fact, if content and function morphemes are characteristically stressed differently, then a simple strategy of attending within phrases to the stressed word would tend to locate phrase heads in a way that could differentiate between the two types of structure. We are currently constructing a connectionist model of this hypothesis differentiation process.

Such exercises are *not* intended as models of the language-acquiring child: Rather, they are demonstrations that certain kinds of information are statistically available in natural discourses, under particular assumptions of input and output representation. Thus, the models exemplify connectionism as a tool to examine the role that associative-activation knowledge could play in language acquisition.

In any event, this application of BC demonstrates how associative structures, learned from degraded and simple input data, can be used to extend limited computational knowledge and capacity: There are many phrase types which children may not be able to compute structurally as phrases; such phrases would still be assigned a correct segmentation by the associatively based habits, trained on evidentially clear cases of where entire utterances actually end. The point of the exercise, again, is *not* to argue that we have demonstrated a connectionist model of phrase assignment, but rather to show that the specific input representations are adequate to learn, from a natural corpus, a set of associations that combine with a boundary sharpening procedure to segment phrases correctly. How this is actually implemented in the child or adult—if it is—is an entirely different matter, for reasons discussed in the first section of this chapter.

CONCLUSION—BEHAVIOURAL STRATEGIES AND MENTAL STRUCTURES

The overall research programme I have outlined suggests a general rehabilitation of the notion of behavioural "strategy." The theoretical claim is that associative strategies form automatically and are natural bases for extending the range of a limited computational capacity. The perceptual strategy that "bigger = more," discussed at the beginning of this chapter, is probably statistically supported over the range of cases that children can compute (roughly, displays with six items or less): hence, the strategy can be derived from computable cases, and then extended to larger cases. The situation is less clear in the case of a complex linguistic strategy, such as (4a), that an "N-V-N" sequence corresponds to "agent, action, object": how can we be sure that this strategy is justified in the language the child hears? One can envisage a model that would be trained to assign basic grammatical relations to a selected set of short motherese sentences, for example, those which use only the set of 100 most frequent words. How well the trained model generalises to new sentences which have the "n-v-n" pattern would be a measure both of the learnability of the pattern on the short sentences, and how well supported it is as a generalisation applying to longer sentences.

Such an investigation will confirm (or disconfirm) the long-held view that actors precede and agents follow their predicates. Other generalisations (e.g. that agents are animate) can be investigated in similar ways. In the general case, one could assign a surface and semantic analysis to a training text, and train a model to assign the basic semantic relations that link the surface phrases: The success of the model at learning the training text will be a measure of the regularities in that text; its success at generalising to new text will be a measure of the availability of statistical information about the surface input to discriminate underlying thematic relations; finally, and perhaps most important for our general research program, analysis of the trained model may reveal *new* informative properties of the surface input.

Connectionist engines offer empirical answers to the problem of analysing statistically available information. What we called perceptual and production "strategies" in language behaviour are interpretable as implementations of statistically valid relations between structural representations. Connectionist models allow us to explore what kinds of such generalisations exist under different representational and environmental assumptions. That is, in the implied long programme of research, the role of connectionist modelling is to inform us about the kinds of statistical information available to the child, given its structural capacities at each stage. This will define the information which children might use to confirm some structural hypotheses and extend the behavioural application of others. Thus, we can use these frequency engines to the hilt to help us understand what the beast *might* be doing to extend the

domains of the demons. But it will be a further empirical matter to see how they co-exist productively in the child.

ACKNOWLEDGEMENTS

A number of the arguments concerning McClelland and Rumelhart were worked out with Joel Lachter. In general, I have learned whatever I know about connectionist modelling from conversations with Lachter, Jerry Feldman, and Gary Dell. Mark Seidenberg and Jay McClelland were generous with their explanations of how their model works, and checked the accuracy of my summary of it. Cornell Juliano is largely responsible for Baby Clauseau.

REFERENCES

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

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

Bever, T.G. (1975a). Functional explanations require independently motivated functional theories. In R. Grossman, J. San, & T. Vance (Eds.), *Papers from the Parasession on Functionalism*, Chicago Linguistic Society. Chicago: University of Chicago, 580–563.

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

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

Bever, T.G. (1987). The aesthetic constraint on cognitive structures. In W. Brand & R. Harrison (Eds.), *The representation of knowledge and belief* (pp. 314–356). Tucson, Arizona: University of Arizona Press.

Bever, T.G., 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, Mass.: M.I.T. Press.

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

Bever, T.G. & Hansen, R. (submitted). *The induction of mental structures while learning to use symbolic systems.*

Bowerman, E. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 319–346). Cambridge, Mass.: Cambridge University Press.

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

Brunswick, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley, C.A.: University of California Press.

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

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

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

Cloitre, M. & Bever, T.G. (1986). *Linguistic anaphors and levels of representation.* Cognitive Science Technical Report No. 36. Rochester, N.Y.: University of Rochester.

Coltheart, M., Sartori, G., & Job. R. (Eds.) (1986). *Cognitive neuropsychology of language.* London: Lawrence Erlbaum Associates Ltd.

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

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.

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

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

Gerken, L.A. (1987). *Function morphemes in young children's speech perception and production.* Columbia University, Ph.D. Dissertation.

Gerken, L.A., Landau, B., & Remez, R. (in press). Function morphemes in young children's speech perception and production. *Developmental Psychology*.

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. I.* Cambridge, Mass.: M.I.T. Press.

Jandreau, S.M., Muncer, S.J., & Bever, T.G. (1986). Improving the readability of text with automatic phrase-sensitive formatting. *British Journal of Educational Technology, 17*, May.

Juliano, C. & Bever, T. (1990). *Clever moms: Regularities in motherese that prove useful in parsing.* Presented at the 1990 CUNY Sentence Processing Conference.

Karmiloff-Smith, A. (1986). Cognitive processes and linguistic representations: Evidence from children's metalinguistic and repair data. *Cognition, 23*, 95–147.

Kuçera, H. & Francis, W.N. (1967). *Computational analysis of present-day American English.* Providence, Rhode Island: Brown University Press.

Lachter, J. & Bever, T.G. (1988). The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models. *Cognition, 28*, 195–247.

Lacoutre, Y. (1989). From mean square error to reaction time: A connectionist model of word recognition. In D. Touretsky & T. Sejnowski (Eds.), *Proceedings of the 1988 connectionist models summer school.* San Mateo, Calif.: Morgan Kauffman.

Langackre, R. (1987). The cognitive perspective. *Reports of the Centre for Research in Language, San Diego, 1*, (3).

Langer, J. (1982). Mental representations and cognitive development. In T.G. Bever (Ed.), *Regression in mental development.* Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

Levelt, W. (in press). *Speaking: From intention to articulation.* Cambridge, Mass.: Bradford Books.

MacWhinney, B. & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*, (121–157).

McClelland, J.L. & Kawamoto, A.H. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J.L. McClelland, D.E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition, Volume 2: Psychological and biological model* (pp. 272–331). Cambridge, Mass.: M.I.T. Press.

Meyer, D.E. & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234.

Minsky, M. & Pappert, S. (1969). *Perceptions.* Cambridge, Mass.: M.I.T. Press.

Newport, E.L. (1977). Motherese: The speech of mothers to young children. In N.J. Castellan, D.B. Pisoni, & G.R. Potts (Eds.), *Cognitive theory, Vol. 2.* Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

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

Plunkett, K. & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition, 38*, 43–102.

Rizzi, L. (1982). *Issues in Italian syntax.* Dordrecht: Foris.

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

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

Rumelhart, D. & McClelland, J. (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, Mass.: M.I.T. 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, 643.

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

Seidenberg, M.S. (1985). Constraining models of word recognition. *Cognition, 20*, 169–190.

Seidenberg, M.S. (1987). Sub-lexical structures in visual word recognition. Access units or orthographic redundancy. In M. Coltheart (Ed.). *Attention and performance XII: Reading* (pp. 245–263). Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

Seidenberg, M.S. (in press). Cognitive neuropsychology and language. *Cognitive Neuropsychology.*

Seidenberg, M.S. (1989). Visual word recognition and pronunciation: A computational model and its implications. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 25–74). Cambridge, Mass.: M.I.T. Press.

Seidenberg, M.S. & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. McGill University Cognitive Science Technical Report. *Psychological Review, 96*, 523–568.

Sejnowski, T.J. & Rosenberg, C. (1986). *NETtalk: A parallel network that learns to read aloud.* (EE and CS Technical Report No. JHU/EECS-86/01). Baltimore, M.D.: Johns Hopkins University.

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: Springer-Verlag.

Slobin, D. & Bever, T.B. (1981). Children use canonical sentence schemas in sentence perception. *Cognition, 12*, 229–265.

St John, M. & McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Valian, V.V. & Coulson, C. (1990). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71–86.

Venezky, R. (1970). *The structure of English orthography.* The Hague: Mouton.

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