Language processing

edited by

Simon Garrod and Martin J. Pickering
University of Glasgow, UK

Psychology Press
a member of the Taylor & Francis group

Tanenhaus, M.K., Spivey, M.J., & Hanna, J.E. (in press). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In M.W. Crocker, C. Clifton, & M. Pickering (Eds.), *Architectures and mechanisms for sentence processing*. Cambridge, UK: Cambridge University Press.

Trueswell, J.C., & Tanenhaus, M.K. (1994). Towards a lexicalist framework of constraint-based syntactic ambiguity resolution. In C.C. Clifton Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 155–180). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Weinberg, A. (1994). Parameters in the theory of sentence processing: Minimal commitment theory goes east. *Journal of Psycholinguistic Research, 22*(3), 339–364.

CHAPTER EIGHT

Connectionism and natural language processing

Nick Chater
University of Warwick, Coventry, UK

Morten H. Christiansen
Southern Illinois University, Carbondale, USA

INTRODUCTION

Many of the chapters of this book are concerned with topics in language processing. This chapter is concerned, by contrast, with a particular method, connectionist computational modelling, which has been applied to a wide range of topics. It is, furthermore, a controversial method: Some have argued that natural language processing from phonology to semantics can be understood in connectionist terms; others have argued that *no* aspects of natural language can be captured by connectionist methods. And the controversy is particularly heated because of the *revisionist* claims of some connectionists: For many, connectionism is not just an additional method for studying language processing, but it offers an alternative to traditional theories, which describe language and language processing in symbolic terms. Indeed, Rumelhart and McClelland (1987, p. 196) suggest "that implicit knowledge of language may be stored among simple processing units organized into networks. While the behaviour of such networks may be describable (at least approximately) as conforming to some system of rules, we suggest that an account of the fine structure of the phenomena of language and language acquisition can best be formulated in models that make reference to the characteristics of the underlying networks." We shall see that the degree to which connectionism supplants, rather than complements, existing approaches to language is itself a matter of debate. Finally, the controversy over connectionist approaches to language is an important test case for the validity of connectionist methods in other areas of psychology.

In the next section, we describe the historical and intellectual roots of connectionism, then introduce the elements of modern connectionism, how it has been applied to natural language processing, and outline some of the theoretical claims that have been made for and against it. We then consider four central topics in connectionist research on language processing: word naming and visual word recognition, lexical processing during speech, morphological processing, and syntax.[1] These illustrate the range of connectionist research on language, give an opportunity to assess its strengths and weaknesses across this range, and allows the general debate concerning the validity of connectionist methods to be illustrated in specific contexts. We would argue that debates in each of these areas, although interrelated, should each be considered on their own merits: It may be that connectionist approaches are valuable in modelling some aspects of language processing, but not in others. Finally, in the Conclusions we sum up and consider the prospects for future connectionist research, and its relation to other approaches to understanding language processing and language structure.[2]

BACKGROUND

From the perspective of modern cognitive science, we tend to see theories of human information processing as borrowing from theories of machine information processing, i.e. from computer science. Within computer science, symbolic processing on general purpose digital computers has proved to be the most successful method of designing practical computational devices. It is therefore not surprising that cognitive science, including the study of language processing, has aimed to model the mind as a symbol processor.

Historically, however, theories of human thought inspired attempts to build computational devices, rather than the other way around. Mainstream computer science arises from the tradition that thought is a matter of symbol processing. This tradition can be traced to Boole's (1854) suggestion that logic and probability theory describe "Laws of Thought", and that reasoning in accordance with these laws can be conducted by following symbolic rules. It runs through Turing's (1936) argument that all human thought can be modelled by symbolic operations on a tape (the Turing machine), through von Neumann's design for the modern digital computer, to the development of symbolic computer programming languages, and thence to modern computer science, artificial intelligence, and symbolic cognitive science.

Connectionism (also known as "parallel distributed processing", "neural networks", or "neurocomputing") can be traced to a different tradition, which attempts to design computers inspired by the structure of the brain. McCulloch and Pitts (1943) provided an early and influential idealisation of neural function. In the 1950s and 1960s, Ashby (1952), Minsky (1954), Rosenblatt (1962), and many others designed various computational schemes based on idealisations of the brain. Aside from their biological origin, these schemes were of interest because they were able to learn from experience, rather than being designed. Such "self-organising" or learning machines therefore seemed prima facie plausible as models of the aspects of human cognition which are learned rather than innate, including many aspects of language processing (although Chomsky, e.g. 1965 was to challenge the extent to which languages are learned). Throughout this period connectionist and symbolic computation stood as alternative paradigms for modelling intelligence, and it was unclear which would prove to be the most successful. But gradually the symbolic paradigm gained ground, resulting in powerful models in the domains such as language (Chomsky, 1957, 1965) and problem solving (Newell & Simon, 1972). The connectionist approach was largely abandoned, particularly in view of the limited power of then current connectionist methods (see, e.g. Minsky & Papert, 1969, for an influential analysis). But some of these limitations have been overcome (Hinton & Sejnowski, 1986; Rumelhart, Hinton, & Williams, 1986), re-opening the possibility that connectionist computation constitutes an alternative to the symbolic model of thought.

So connectionism is inspired by the structure and processing of the brain. What does this mean in practice? At a coarse level of analysis, the brain can be viewed as consisting of a very large number of simple processors, neurons, which are densely interconnected into a complex network; and these neurons do not appear to tackle information processing problems alone—rather, large numbers of neurons operate co-operatively, and simultaneously, to process information. Furthermore, neurons appear to communicate numerical values (encoded by firing rate), rather than passing symbolic messages, and, to a first approximation at least, neurons can be viewed as mapping a set of numerical inputs (delivered

[1] It should be noted that many connectionist models cut across this traditional division into different aspects of language. Thus, such a division may perhaps do injustice to connectionist models of language (Sharkey, 1991)—or even lead into an "incommensurability trap" (Christiansen & Chater, 1992). It is, however, merely meant to reflect the topics addressed in the rest of this book.

[2] For a survey of current research on connectionist natural language processing, see the Special Issue of the journal *Cognitive Science*, "Connectionist models of language processing: progress and prospects" (Christiansen, Chater, & Seidenberg, in press).

from other neurons) onto a numerical output (which is then transmitted to other neurons). Connectionist models are designed to mimic these properties: Hence, they consist of large numbers of simple processors, known as *units* (or nodes), which are densely interconnected into a complex network, and which operate simultaneously and co-operatively to solve information processing problems. In line with the assumption that real neurons are numerical processors, units are assumed to pass only numerical values rather than symbolic messages, and the output of a unit is usually assumed to be a numerical function of its inputs. Typical connectionist networks do not amount to realistic models of the brain, however (see, e.g. Sejnowski, 1986), either at the level of the individual processing unit, which not only drastically oversimplifies, but knowingly falsifies, many aspects of the function of real neurons, or in terms of the structure of the neural networks, which bear little if any relation to brain architecture. One avenue of research is to seek increasing biological realism (e.g. Koch & Segev, 1989). In the study of aspects of cognition in which little biological constraint is available, most notably language, researchers have concentrated on developing connectionist models with the goal of accurately modelling human behaviour. They therefore take their data from cognitive psychology, linguistics, and cognitive neuropsychology, rather than from neuroscience. Here, they must compete head-on with symbolic models of language processing.

We noted earlier that the relative merits of connectionist and symbolic models of language are hotly debated. But should they be viewed as standing in competition at all? Advocates of symbolic models of language processing assume that symbolic processes are somehow implemented in the brain. Thus, they too are connectionists, at the level of *implementation*. They assume that language processing can be described at two levels: at the psychological level, in terms of symbol processing; and at the implementational level, in neuroscientific terms (to which connectionism approximates). If this is right, then connectionist modelling should proceed by taking symbol processing models of language processing, and attempting to implement these in connectionist networks. Advocates of this view (Fodor & Pylyshyn, 1988; Pinker & Prince, 1988) typically assume that it implies that symbolic modelling should be entirely autonomous from connectionism; symbolic theories set the goalposts for connectionism, but not the other way round. Chater and Oaksford (1990) have argued that, even according to this view, there will be two-way influence between symbolic and connectionist theories, since many symbolic accounts can be ruled out precisely because they could not be neurally implemented. But most connectionists in the field of language processing have a more radical agenda: not to implement, but to challenge, to varying degrees, the symbolic approach to language processing.

Before outlining and evaluating a range of specific connectionist models of language processing, it is useful to set out some of the recurring themes in discussion of the virtues and vices of the connectionist approach to language:

Learning. As discussed previously, connectionist networks typically, although not always, learn from experience,[3] rather than being fully specified by a designer. Symbolic computational systems, including those concerned with language processing, are typically, but not always, fully specified by the designer.

Generalisation. Few aspects of language are simple enough to be learnable by rote. The ability of networks to generalise to cases on which they have not been trained is thus a critical test for many connectionist models.

Representation. Because they are able to learn, the internal codes used by connectionist networks need not be fully specified by a designer, but are devised by the network so as to be appropriate for the task. Developing methods for understanding the codes that the network develops is an important strand of connectionist research. Whereas internal codes may be learned, the inputs and outputs to a network generally use a code specified by the designer. These codes can be crucial in determining network performance, as we shall see. How these codes relate to standard symbolic representations of language in linguistics is a major point of contention.

Rules versus exceptions. Many aspects of language can be described in terms of what have been termed "quasi-regularities"—regularities that are usually true, but which admit some exceptions. According to the symbolic descriptions used by modern linguistics, these quasi-regularities may be captured in terms of a set of symbolic rules, and sets of exceptions to those rules. Processing models often incorporate this distinction by having separate mechanisms to deal with rule-governed and exceptional cases. It has been argued that connectionist models provide a single mechanism, which can pick up general rules, while learning the exceptions to those rules. Although this issue has been, as we shall see, a major point of controversy surrounding connectionist models, it is important to note that attempting to provide single mechanisms for rules and exceptions is not

[3] Although important in many connectionist models, we will not provide a detailed account of connectionist approaches to language *acquisition* here. For an overview, see Plunkett (1995) and for discussion of possible consequences for traditional approaches to language acquisition, see Seidenberg (1994).

essential to the connectionist approach; one or both separate mechanisms for rules and exceptions could themselves be modelled in connectionist terms (Coltheart, Curtis, Atkins, & Haller, 1993; Pinker, 1991; Pinker & Prince, 1988). A further question is whether networks really learn rules at all, or whether they simply approximate rule-like behaviour. Opinions differ concerning whether the latter is an important positive proposal, which may lead to a revision of the role of rules in linguistics (Rumelhart & McClelland, 1986a; see also Smolensky, 1988), or whether it is a fatal problem with connectionist models of language processing (Pinker & Prince, 1988).

With these general issues in mind, let us consider some of the broad spectrum of connectionist models of language processing.

VISUAL WORD RECOGNITION AND WORD NAMING

The psychological processes engaged in reading are extremely varied and complex, ranging from early visual processing of the printed word, to syntactic, semantic, and pragmatic analysis, to integration with general knowledge. Connectionist models have concentrated on very simple aspects of the reading process: (1) recognising words from printed text, and (2) word "naming", i.e. mapping visually presented letter strings onto sequences of sounds (this may or may not involve word recognition). We focus on connectionist models of these two processes here.

One of the earliest connectionist models was McClelland and Rumelhart's "interactive activation" (1981) model of visual word recognition (see also Rumelhart & McClelland, 1982). The network is completely prespecified (i.e. it does not learn), and consists of a sequence of "layers" of units, as illustrated in Fig. 8.1. Units in the first layer are specific to particular visual *features* of letters (in particular positions within the word). Units in the second layer stand for particular letters (also in particular positions within the word). Units in the third layer stand for words. Within and between layers, there are inhibitory connections between units which stand for incompatible states of affairs. For example, there are inhibitory connections between units in the word layer, so that possible "candidate" words compete against each other. There are also excitatory connections between units which stand for mutually reinforcing states of affairs at different layers. For example, there is an excitatory connection between the unit standing for the word TAKE, the unit standing for the letter "T" (in the first position) as well as the particular letter features which make up "T". All excitatory connections of a given kind—for example, between the letter level and the word level units—have the same strength, but this strength varies depending on which two levels are involved. This is also the case with the inhibitory connections between and within unit levels. This defines the "architecture" of the network (shown in Fig. 8.1).

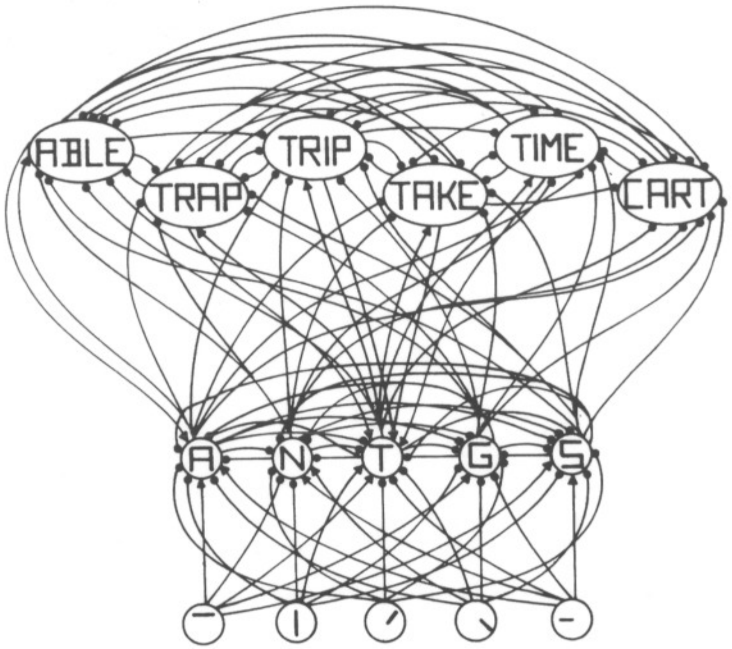

FIG. 8.1 The letter "T" in the first position of a word consisting of four letters, some of its neighbouring nodes, and their interconnections in the McClelland and Rumelhart (1981) interactive activation model of visual word recognition. Excitatory connections are shown as arrows, whereas inhibitory connections have circular ends. From "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings" by J.L. McClelland and D.E. Rumelhart, 1981, *Psychological Review*, *88*, p. 380. Copyright © (1981) by the American Psychological Association. Reprinted with permission

How do individual units behave? In interactive activation models such as this, the level of activity of a unit is determined by its previous level and its current input (as we shall see below, in more recent models, the state of a unit is typically determined only by its current input). If the input to a unit is 0, then all that happens is that the level of activity of the unit decays exponentially. The input to the unit is, as is standard, simply the weighted sum of the units which are inputs to that unit (where the weights correspond to the strengths of the connections). If the input is positive, then the level of activity is increased in proportion both to that input, and to the distance between the current level of activation and the maximum activation (conventionally set at 1); if the input is negative, the level of activity is decreased in proportion to the input, and to the distance between the current level of activation and the minimum activation (conventionally set at –1, but in the McClelland & Rumelhart, 1981, model it was set at –0.2 to allow rapid reactivation).

Although this behaviour sounds rather complex, the basic idea is simple. Given a constant input, the unit will gradually adjust to a stable level where the exponential decay balances with the boost from that input: Positive constant inputs will be associated with positive stable activation, negative constant inputs with negative stable activation; and small inputs lead to activations levels close to 0, whereas large inputs lead to activation values which tend to be near 1 or –1. An activation level near 1 corresponds to a high level of confidence that an item is present; an activation level near –1 corresponds to a high level of confidence that it is not.

Word recognition occurs as follows. A visual stimulus is presented, which activates in a probabilistic fashion the units in the first layer, standing for visual features. Depending on the particular experimental task being modelled (e.g. recognising a bright, high-contrast target followed by mask, or a degraded target), the probability of a feature being activated is set to 1.0 or below. As the features become activated, they send activation via their excitatory and inhibitory connections to the units at the letter level. Notice that so far only bottom-up flow of information has taken place—there is no inhibition between the feature units and no feedback from the letter level to the feature level. In addition, the weights of the inhibitory connections between the letter units (shown in Fig. 8.1) were set to 0, meaning that no inhibition takes place between the letters.[4] As the letter units become activated they, in turn, send excitatory and inhibitory activation to the word-level units. The words compete amongst each other via their inhibitory connections, and reinforce their component letter units via excitatory feedback to the letter level (there is no word-to-letter inhibition). At this point, an "interactive" process is thus occurring between the letter level and the word level: Bottom-up flow of information from the visual input is combined via the activation of the letter units with the top-down information flow from the word units. The entire process involves a cascade of overlapping and interacting processes: letter and word recognition do not occur one after the other as distinct processing stages, but rather are mutually constraining.

[4] The theoretical model, which motivated the simulation model described here, is meant to be fully "interactive", with mutual inhibition between competing units at the same level as well as bi-directional excitatory and inhibitory connections between the three levels, but this was not implemented.

The interactive character of McClelland and Rumelhart's model embodies a controversial theoretical claim about reading. Many researchers have assumed that reading involves the successive computation of increasingly abstract levels of representation, but that there is no feedback from more abstract to less abstract levels. This kind of account is sometimes known as "bottom-up" and can also be realised in connectionist networks, as we shall see later. The question of whether reading is bottom-up or interactive has been a major focus of debate. We shall see later that the same debate rages in the speech perception literature; and analogous issues arise throughout perception (e.g. Bruner, 1957; Fodor, 1983; Marr, 1982; Neisser, 1967).

This model proved able to account for a variety of phenomena, mainly concerning contextual effects on perception of single letters. For example, it captures the fact that letters presented in the context of a word are recognised more rapidly than letters presented individually, or in random letter strings (Johnston & McClelland, 1973). This is because the activation of the word containing a particular letter provides top-down confirmation of the identity of that letter, in addition to the activation provided by the bottom-up feature-level input. Moreover, it has been shown that letters presented in the context of pronounceable non-words (i.e. pseudowords, such as "mave", which are consistent with English phonotactics) are recognised more rapidly than letters presented singly (Aderman & Smith, 1971) or in contexts of random letter strings (McClelland & Johnston, 1977). In this case, the facilitation is caused by a "conspiracy" of partially activated similar words, which are triggered in the non-word context, but not in the random letter string context. These partially active words provide a top-down confirmation of the letter identity, and thus they "conspire" to enhance recognition. In a similar fashion, the model explains how degraded letters can be disambiguated by their letter context, and how occurring in a word context can facilitate the disambiguation of component letters even when they are all visually ambiguous. Moreover, it provides an impressively detailed demonstration of how interactive processing can account for a range of further experimental effects. As we shall see later, however, not all theorists agree that interactive processes are required to explain these and other phenomena in language processing.

Recent work on connectionist modelling of reading has had a somewhat different focus: on word naming rather than recognition. It has been concerned with the problem of learning the relationship between written word forms and their pronunciations, although, as we shall see, issues of word recognition also arise. The first such model was Sejnowski and Rosenberg's (1987) NETtalk—shown in Fig. 8.2—which learns to read aloud from written text.

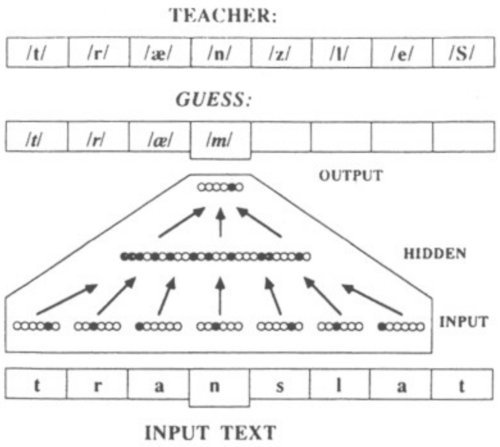

FIG. 8.2. Illustration of the NETtalk architecture. The input layer consists of 203 units divided into 7 groups, which each correspond to a particular letter position. Information from the inputs is fed forward via 80 hidden units to an output layer containing 26 units. During training the network's guess (the activation of the output units) is compared with the desired target provided by a teacher, and network weights are then subsequently altered so as to minimise any discrepancy. From "Neural representation and neural computation" by P.S. Churchland and T.J. Sejnowski in *Neural Connections, Mental Computations*, edited by L. Nadel, L. Cooper, P. Culicover and R.M. Harnish, 1989, MIT Press. Copyright © (1989) MIT Press. Reprinted with permission.

NETtalk uses a *feedforward*, rather than an interactive, network architecture. In a feedforward network, the units are, as before, divided into layers, but activation flows only in one direction through the network, starting at the layer of "input units", and finishing at the layer of "output units". The internal layers of the network are known as "hidden units". There may be several hidden layers in a feedforward network, but in NETtalk, as in many neural networks, there is just one. The input units represent a "window" of consecutive letters of text. The output units represent the network's suggested pronunciation for the middle letter. The network can be used to pronounce a written text by shifting the moving window across the text, letter by letter, so that the central letter to be pronounced moves onwards a letter at a time. In English orthography, there is not, of course, a one-to-one mapping between letters and phonemes. NETtalk uses a rather ad hoc strategy to deal with this: in clusters of letters realised as a single speech sound (e.g. "th", "sh", "ough") only one of the letters is chosen to be mapped onto the speech sound, and the others are not mapped onto any speech sound.

The behaviour of individual units is rather simpler than in the interactive activation network. The activation of each unit is determined by its current input (calculated as the weighted sum of its inputs, as before): Specifically, this input is "squashed", so that the activation of each unit lies between 0 and 1. As the input to a unit tends to positive infinity, the level of activation approaches 1; as the input tends to negative infinity, the level of activation approaches 0. With occasional minor variations, this description applies equally to almost all feedforward connectionist networks.

Whereas the interactive activation model was prespecified, NETtalk learns from exposure to text associated with the correct pronunciation. Specifically, the network is presented with inputs representing seven letter contexts through English texts; and with each input, it is given the "target" output, i.e. the output which corresponds to the correct pronunciation. The inputs use a "position-specific", letter level representation, i.e. the input units are divided into discrete banks, each corresponding to one of the seven letter positions, and within each bank, units correspond to specific letters. At the beginning of training, NETtalk's output bears no relation to the correct pronunciation; but after extensive training, its standard of pronunciation is good enough to be largely comprehensible when fed through a speech synthesizer.

How is learning achieved? Like many of the connectionist models we shall describe later, NETtalk is trained by "back-propagation" (Rumelhart et al., 1986, prefigured in Bryson & Ho, 1975; Werbos, 1974). When each input is presented, it is fed through the network, and the output is derived. The output is compared against the correct "target" value and the difference between the two is calculated for each output unit. The squared differences are summed over all the output units, to give an overall measure of the "error" that the network has made. The goal of learning is to reduce overall level of error, averaged across input/target pairs (in this context, this means averaging across typical texts). Back-propagation is a procedure which specifies how the weights of the network (i.e. the strengths of the connections between the units) should be adjusted in order to decrease the error. Training with back-propagation is guaranteed (within certain limits) to reduce the error made by the network. If everything works well, then the final level of error may be very small, meaning that the network produces the desired output. Notice that the network will produce an output not only for inputs on which it has been trained, but for any input. If the network has learned about regularities in the mapping between inputs and targets, then it should be able to *generalise* successfully to new items. NETtalk is able to pronounce letters in contexts that it has never before encountered reasonably successfully.

Back-propagation may sound too good to be true.[5] But note that back-propagation merely guarantees to adjust the weights of the network to *reduce* the error; it does not guarantee to reduce the error to 0, or a value anywhere near 0. Indeed, in practice, back-propagation can configure the network so that error is very high, but changes in weights in any direction lead to the same or a higher error level. This is known as the problem of local minima. Attempting to avoid this problem is a major day-to-day concern of connectionist researchers, as well as being a focus of theoretical research. Local minima can be avoided by judicious choice among the large number of variants of back-propagation, and by appropriate decisions on the numerous parameters involved in model building (such as the number of hidden units used, whether learning proceeds in small or large steps, and many more). Despite these problems, back-propagation is surprisingly successful in many contexts. Indeed, the feasibility of back-propagation learning has been one of the reasons for the renewed interest in connectionist research. Prior to the discovery of back-propagation, there were no well-justified methods for training multilayered networks. The restriction to single-layered networks was unattractive, since Minsky and Papert (1969) showed that such networks, sometimes known as "Perceptrons" have very limited computational power. It is partly for this reason that hidden units are viewed as having such central importance in many connectionist models; without hidden units, most interesting connectionist computation would not be possible.

What internal code on the hidden units is NETtalk using? This code is not prespecified by the designer, but is learned from experience by the network. Furthermore, it turns out that the pattern of hidden units does not have a transparent interpretation to the casual observer. Sejnowski and Rosenberg gained some insight into what their network is doing by first computing the average hidden unit activation given each of a total of 79 different letter-to-sound combinations. For example, the activation of the hidden unit layer was averaged for all the words in which the letter "c" is pronounced as /k/, another average calculated for words in which "c" corresponds to /s/, and so on. Next, the relationships among the resulting 79 vectors—each construed as the network's internal representation of a particular letter-to-sound correspondence—were explored via cluster analysis. Interestingly, all the vectors for vowel sounds clustered together, suggesting that the net had learned to treat vowels as different from consonants. Moreover, the net had learned a number of sub-regularities amongst the letter-to-sound combinations, evidenced for example by the close clustering of the labial stops /p/ and /b/ in hidden unit space.

NETtalk was intended as a demonstration of the power of neural networks, rather than as a psychological model. Seidenberg and McClelland (1989) provided the first detailed psychological model of reading aloud. They also used a feedforward network with a single hidden layer, but they represented the entire written form of the word as input, and the entire phonological form as output. This network implemented one side of a theoretical "triangle" model of reading in which the two other sides were a pathway from orthography to semantics and a pathway from phonology to semantics (these sides are meant to be bi-directional and, in fact, the implemented network also produced a copy of the input as a second output to attempt to model performance on lexical decision tasks, but we shall ignore this aspect of the model here). Seidenberg and McClelland restricted their attention to 2897 monosyllabic words of English, rather than attempting to deal with unrestricted text like NETtalk.

The orthographic and phonological representations used by Seidenberg and McClelland are rather complex, and we give just a sketch here. The most straightforward style of representation would be to use position-specific codes for each letter or phoneme. But this seems unattractive, partly because it fails to capture the fact that the mapping between letters and sounds is (roughly) the same wherever those letters or sounds occur on the word. Using a position-specific code, the network must learn afresh that the letter "t" often maps onto the phoneme /t/ for every position. This is because letters and sounds are represented by distinct units in each position. Indeed, the position-specific scheme does not seem just unattractive—it makes the absurd prediction that an orthographic system in which the correspondence between letters and sounds was different for every serial position should present no special problems. The network would learn this kind of "scrambled" mapping just as easily as normal English orthography; but the human learner would presumably be dramatically impaired. Another difficulty with position-specific encodings is that since, as discussed earlier, letters and phonemes do not stand in one-to-one correspondence, the network would have to solve a difficult "alignment" problem. As we noted, NETtalk finesses this problem by the designer prespecifying a particular alignment; we shall see later that it is also possible for the network itself to solve the alignment problems using a NETtalk style of position-specific representation (Bullinaria, 1994). But Seidenberg and McClelland sidestep the problem by using an ingenious strategy.

[5] In fact, it is most likely not a biologically plausible learning algorithm. Still, back-propagation provides a convenient learning method which may result in networks with computational properties similar to those of real neural structures.

The idea is to decompose both the letter and phoneme strings into consecutive triples. Thus, the letter string FISH is decomposed into _FI, SH_, ISH, FIS. Notice that the triples are position-independent, but that the overall string can be pieced together again from the triples (in general, as Pinker & Prince, 1988 have noted, this piecing-together process cannot always be carried out successfully, but in this context it is adequate). The phonemic string is also decomposed into triples of phonemes. Rather than represent the phonemes directly, units are devoted to triples of features of phonemes. This style of representation, termed wickelfeatures (after Wickelgren, 1969, who employed triples in modelling memory for sequential material), was first used in Rumelhart and McClelland's (1986a) model of learning the English past tense, which we will discuss later in the section on morphological processing. In the orthographic layer, each unit is associated with a list of 1000 random letter triplets (10 possible first letters × 10 possible middle letters × 10 possible end letters) and is activated if one of the letter triplets in the input occurs in this list.

Seidenberg and McClelland trained their network to produce an output corresponding to wickel-representation of the pronunciation of a word, from a wickel-representation of its orthography given as input. The performance of the network captures a wide range of experimental data (on the reasonable assumption that network error can be roughly equated with response time in experimental paradigms). For example, frequent words are read more rapidly (with lower error) than rare words (Forster & Chambers, 1973); orthographically regular words are read more rapidly than irregulars, and the difference between regulars and irregulars is much greater on rare rather than frequent words (Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Taraban & McClelland, 1987).

Seidenberg and McClelland's model uses a single mechanism to capture both the rules governing the pronunciation of English text and the exceptions to those rules. This contrasts with the standard view of reading, according to which the rules and the exceptions are treated separately. Indeed, it is standard to assume that there are two distinct routes in reading, a so-called "phonological route", which applies rules of pronunciation, and a so-called "lexical route", which is simply a list of words and their pronunciations. The idea is that regular words can be read using either route; that irregulars must be read by using the lexical route, to override the phonological route; and that non-words can be pronounced by using the phonological route (these will not be mentioned in the lexical route). Seidenberg and McClelland claim to have shown that this *dual-route* view is not necessarily correct, since a single route can pronounce both irregular words and non-words. Furthermore, they have provided a fully explicit computational model, whereas dual-route theorists have merely sketched the reading system at the level of "boxes and arrows" (though see Coltheart, Curtis, Atkins & Haller, 1993 for a recent exception).

A number of criticisms have been levelled at Seidenberg and McClelland's account, however, and we briefly consider some of these. First, can a single route really account for both non-word and exception word pronunciation? Besner, Twilley, McCann, and Seergobin (1990) have argued that the non-word reading performance of Seidenberg and McClelland's model is actually very poor compared with human readers (though see Seidenberg & McClelland, 1990 for a reply). Moreover, Coltheart et al. (1993) have argued that better performance at non-word reading can be achieved by symbolic learning methods, using the same word-set as Seidenberg and McClelland.

Another limitation of the Seidenberg and McClelland model is the use of frequency compression during training. Rather than present rare and frequent words equally often to the network, they presented words with a probability proportional to their log frequency of occurrence in English (using Kucera & Francis, 1967). Had they used raw frequency, rather than log frequency, the network could have encountered low frequency items too rarely to learn them at all; this must be counted as a difficulty for this and many other network models, since the human learner must deal with absolute frequencies. Recently, however, Plaut, McClelland, Seidenberg, and Patterson (1996) have demonstrated that a feedforward network can be trained successfully using the *actual* frequencies of words instead of their log frequency[6]—even to a level of performance similar to that of human subjects on both word and non-word pronunciation.

At a more technical level, Seidenberg and McClelland's model is limited in that it does not readily extend to deal with words with more than one syllable. Furthermore, the use of wickelfeatures creates a number of problems. One of the most important is that output from the network cannot readily be interpreted: There is no straightforward decoding from a muddle of partially activated output units representing wickelfeatures to a pronunciation, specified in a standard phonological format (or any sequential format that would seem to be required to drive speech). This meant that Seidenberg and McClelland had to assess the pronunciation intended by their network by considering various plausible pronunciations, converting these into wickelfeatures, and seeing which is the closest to the performance of the model. A better output code would code pronunciation explicitly, rather than burying it in a deeply encrypted form.

[6] Note that Plaut et al. (1996) used these (actual) frequencies to scale the contribution of error for each word during back-propagation training, rather than to determine the number of word presentations. As mentioned later, they also employed a different representational scheme (due to Plaut & McClelland, 1993) than Seidenberg and McClelland (1989).

Recently connectionist work on reading has attempted to take account of these difficulties. For example, Plaut and McClelland (1993) abandon wickelfeatures, and use a localist code, which is loosely position-specific, but which exploits some regularities in English orthography and phonology to avoid using a completely position-specific representation. This learns to read non-words very well—at levels comparable with human non-word reading. But it does so by building in a lot of knowledge into the representation, rather than having the network pick up this knowledge. One could plausibly assume (cf. Plaut et al., 1996) that this knowledge is acquired prior to reading acquisition; that is, children normally know how to pronounce words (i.e. talk) before they start learning to read. This hypothesis was tested by Harm, Altmann & Seidenberg (1994) who demonstrated how pretraining a network on phonology can facilitate the subsequent acquisition of a mapping from orthography to phonology.

One of the problems with this novel representational scheme is, however, that it only works for monosyllabic words. Bullinaria (1994), on the other hand, also obtains very high non-word reading performance, which applies to words of any length. To do so, he gives up the attempt to provide a single route model of reading, and aims only to model the phonological route: He uses a variant of NETtalk, in which orthographic and phonological forms are not prealigned by the designer. The rough idea is that, instead of having a single output pattern, the network has many output patterns corresponding to all possible alignments of the phonology with the orthography. All of these possibilities are considered, and the one that is nearest to the network's actual output is taken to be the correct output pattern, and used to adjust the weights. This approach, like NETtalk, uses an input window which moves gradually over the text, producing one phoneme at a time. Hence, a simple phoneme-specific code can be used; the order of the phonemes is implicit in the order in which the network produces them.

A further criticism of Seidenberg and McClelland's single-route model is that it does not appear to account for an apparent double dissociation between phonological and lexical reading in neuropsychological patients. On the one hand, surface dyslexics (e.g. Marshall & Newcombe, 1973) can read exception words, but not non-words; on the other, phonological dyslexics (e.g. Funnell, 1983) can pronounce non-words but not irregular words. The standard inference from double dissociation to modularity of function (e.g. Shallice, 1988) suggests that normal non-word and exception word reading are subserved by distinct systems—that is, to a dual-route model (Coltheart, 1985; Morton & Patterson, 1980)—although it is important to keep in mind that such double dissociations are never clear-cut. Acquired dyslexia can be simulated by damaging Seidenberg and McClelland's network in various ways (e.g. removing connections or units); although the results of this damage do have neuropsychological interest (Patterson, Seidenberg, & McClelland, 1989), they do not give rise to the double dissociation: an analogue of surface dyslexia is found (i.e. regulars are preserved), but no analogue of phonological dyslexia is observed. Furthermore, Bullinaria and Chater (1995) have explored a range of rule-exception tasks using feedforward networks trained by back-propagation, and concluded that, although double dissociations do occur with single-route models, this only occurs with very small-scale networks. With large networks, the dissociation in which the rules are damaged but the exceptions are preserved does not occur. It remains possible that some realistic single-route model of reading, incorporating factors that have been claimed to be important to connectionist accounts of reading such as word frequency and phonological consistency effects (cf. Plaut et al., 1996) might give rise to the relevant double dissociation.[7] However, Bullinaria and Chater's results indicate that modelling phonological dyslexia is potentially a major difficulty for any single-route connectionist model of reading. Perhaps for this reason, some of the most recent connectionist models of reading now implement an additional "semantic" route.[8]

Single- and dual-route theorists argue about whether non-word and exception word reading is carried out by a single system, but both believe in an additional semantic route for reading. In this route pronunciation is retrieved through accessing a semantic code from the orthographic form. The availability of this additional semantic pathway is evidenced by deep dyslexics, who make semantic errors in reading aloud, such as reading the word *peach* aloud as "apricot". Plaut et al. (1996) argue that this route also plays a role in normal reading. In particular, they suggest that a division of labour emerges between the phonological and the semantic pathway during reading acquisition: Roughly speaking, the phonological pathway moves towards a specialisation in regular (consistent) orthography-to-phonology mappings at the expense of exception words which become the main focus of the semantic pathway.

[7] Whereas "regularity" (the focus of the Bullinaria & Chater simulations) can be taken as indicating that the pronunciation of a word appears to follow a rule, "consistency" refers to how well a particular word's pronunciation agrees with other similarly spelled words. The magnitude of the latter depends on how many "friends" a word has (i.e. the summed frequency of words with similar spelling patterns and similar pronunciation) compared with how many "enemies" (i.e. the summed frequency of words with similar spelling patterns but different pronunciations) (Jared, McRae, & Seidenberg, 1990).

[8] Recall that the theoretical model, motivating the original Seidenberg and McClelland (1989) simulation model, included additional pathways from orthography to semantics and from phonology to semantics, but these were not implemented.

The putative effect of the latter pathway was simulated by Plaut et al. as extra input to the phoneme units in a feedforward network trained to map orthography to phonology. The strength of this external input is frequency dependent and gradually increases as learning progresses. As a result the network comes to rely on this extra phonological input. If eliminated (following a lesion to the semantic pathway), the network loses much of its ability to read exception words, but retains good reading of regular words as well as non-words. In this way, Plaut et al. provides a more accurate account of surface dyslexia than Patterson et al. (1989). In contrast, if the phonological pathway is selectively damaged the resulting deficit pattern should resemble that of phonological dyslexia: reasonable word reading but impaired non-word reading—but this hypothesis was not tested directly by Plaut et al.

Furthermore, the theoretical triangle model of Seidenberg and McClelland (1989)—as implemented in a recent (toy) model (Seidenberg & Harm, 1995)—offers an additional explanation of phonological dyslexia, but in the context of development, rather than as an acquired disorder. This first implementation of the full triangle model employs a so-called *recurrent* network (which, broadly speaking, is akin to a feedforward network, except that units in a particular layer are able to feedback onto units at the same layer).[9] The network thus implements the two "connectionist" routes to reading: either via the orthography–semantics pathway or via the orthography–phonology–semantics pathway. In this model, selective damage to the recurrent feedback connections in the phonological layer may provide an alternative explanation of phonological dyslexia. According to this view, in some kinds of development, and perhaps also acquired phonological dyslexia, *dyslexia* may (in some cases) simply be a misnomer—patients should encounter difficulty with *repeating* non-words, just as much as reading them.[10] Unfortunately, this hypothetical explanation has not been explored in simulations. So, while "lesioned" connectionist networks have been shown to model surface dyslexia quite successfully, no explicit simulations have been presented testing the connectionist explanations of phonological dyslexia.

We have considered connectionist models of reading in some detail, since they introduce the principal connectionist methods, and some of the key debates surrounding connectionist models. We have seen that a range of connectionist accounts have provided a good fit with much of the data on normal and impaired reading, although points of controversy remain. Moreover, connectionist models have contributed to re-evaluation of core theoretical issues, such as whether reading is interactive or purely bottom-up, and whether rules and exceptions are dealt with separately or by a single cognitive mechanism. In subsequent sections we shall see these issues, and others that arise in models of reading, are also important sources of debate concerning connectionist models of other areas of language processing.

[9] That is, typically there is no feedback from higher to lower layers, as in an interactive architecture, but simply connections allowed within a layer (although there are a few specialised recurrent learning algorithms, e.g. Pearlmutter, 1989, allowing feedback connections between layers). A simple variant of these recurrent networks is discussed further in the next section.

[10] The empirical data concerning repetition abilities of putative dyslexics is highly controversial.

LEXICAL PROCESSING DURING SPEECH

Just as connectionist models in reading use two principal architectures, interactive activation, and feedforward networks trained by back-propagation, so with connectionist models of lexical processing during speech.

Speech perception

Interactive activation networks have been used to model both speech recognition and production. The early and very influential TRACE model of speech perception (McClelland & Elman, 1986) consists of a standard interactive activation architecture with layers of units standing for phonetic features, phonemes and words. There are several copies of each layer of units, standing for different points in time in the utterance, and the number of copies differs for each layer. At the featural level, there is a copy for each discrete "time slice" into which the speech input is divided. At the phoneme level, there is a copy of the detector for each phoneme centred over every three time slices. The phoneme detector centred on a given time slice is connected to feature detectors for that time slice, and also to the feature detectors for the previous three and subsequent three slices. Hence, successive detectors for the same phoneme overlap in the feature units with which they interact. Finally, at the word level, there is a copy of each word unit at every three time slices. The window of phonemes with which the word interacts corresponds to the entire length of the word. Here, again, adjacent detectors for the same word will overlap in the lower level units to which they are connected. In short, then, we have a standard interactive activation architecture, with an additional temporal dimension added, to account for the temporal character of speech input.

The debate between interactive and bottom-up models of speech perception parallels the debate between interactive and bottom-up accounts of reading. McClelland and Elman have two kinds of argument in favour of their position. First, and perhaps most important, is the broad coverage of the model in accounting for a range of empirical data on speech perception. For example, TRACE's interactive architecture nicely accounts for the apparent influence of lexical context on phoneme identification. Specifically, TRACE models Ganong's (1980) demonstration that the identification of a syllable-initial speech sound that was constructed to be between a /g/ and a /k/ was influenced by whether the rest of the syllable ended "iss" (making giss or kiss) or "ift" (making *gift* or *kift*). Specifically, the identification of the intermediate phoneme was biased towards the choice that completed a word rather than a non-word. This effect is particularly interesting since the identification of a phoneme appears to be affected by *subsequent* material. (Notice that this phenomenon is directly analogous to the facilitation of letter recognition in word or word-like contexts, discussed earlier.) TRACE captures this effect because phoneme and lexical identification occur in parallel and are mutually constraining. TRACE also captures experimental findings concerning various factors affecting the strength of the lexical influence (e.g. Fox, 1984), and aspects of the categorical aspects of phoneme perception (Massaro, 1981; Pisoni & Tash, 1974). TRACE also provides rich predictions concerning the time-course of spoken word recognition (e.g. Cole & Jakimik, 1978; Marslen-Wilson, 1973; Marslen-Wilson & Tyler, 1975), and lexical influences on the segmentation of speech into words (e.g. Cole & Jakimik, 1980). Although TRACE does account for an impressive range of phenomena, as in the case of reading, bottom-up connectionist models have been proposed which aim to account for a similar range of data (e.g. Norris, 1993; Shillcock, Lindsey, Levy, & Chater, 1992).

Second, McClelland and Elman can derive specific empirical predictions from their model which appear to be incompatible with any bottom-up model. Elman and McClelland (1988) conduct what is intended to be a crucial experiment between interactive and bottom-up approaches, and find the interactive view to be confirmed. The central theoretical question at issue is whether or not lexical effects on phoneme restoration are caused, as the interactive view supposes, by the feedback of information from the lexical to the phonemic level. At first glance, it might appear that these lexical effects simply directly demonstrate that this top-down feedback does occur. But there is an alternative explanation, which is entirely compatible with the modular view: that subjects' decisions concerning which phoneme was heard are influenced by both phonological and lexical representations of the stimulus. According to this view, the lexical level directly influences the subject's decision, without any top-down influence on the phoneme detection process itself.

Experimentally disentangling these two explanations is extremely difficult. But Elman and McClelland noticed a prediction of TRACE which appeared to suggest an appropriate crucial experiment. In natural speech, the pronunciation of a phoneme will to some extent be altered by the phonemes that surround it, in part for articulatory reasons: this phenomenon is known as coarticulation. This means that listeners should adjust their category boundaries depending on the phonemic context. Experiments confirm that people do indeed exhibit this "compensation for coarticulation" (Mann & Repp, 1980). For example, given a series of synthetically produced tokens between /t/ and /k/, listeners move the category boundary towards the /t/ following a /s/ and towards the /k/ following a /sh/. This phenomenon suggests a way of detecting whether lexical information really does feed back to the phoneme level. Elman and McClelland considered the case where compensation for coarticulation occurs across word boundaries, for example, a word-final /s/ influencing a word-initial /t/ as in *Christmas tapes*. If lexical-level representations feed back on to phoneme-level representations, the compensation of the /t/ should still occur when the /s/ relies on lexically driven phoneme restoration for its identity (i.e. in an experimental condition in which the identity of /s/ in *Christmas* is obscured, the /s/ should be restored and thus compensation for coarticulation proceeds as normal). Elman and McClelland noticed that the TRACE model does indeed produce this prediction; and that it is difficult to see how a modular account of speech perception could make the same prediction. They therefore decided to conduct the crucial experiment.

Subjects heard pairs of words such as *Christmas tapes* or *foolish capes*, where the last segment of *Christmas* or *foolish* was replaced by a synthetic segment midway between /s/ and /sh/. The first segment of *tapes*/*capes* was a synthetic segment drawn between /t/ and /k/. Subjects were required to report the identity of the second word. Their responses revealed that the restored identity of the ambiguous phoneme at the end of the first word affected the identification of the ambiguous phoneme at the beginning of the second word in a way which paralleled the compensation effect when unambiguous phonemes were present. Elman and McClelland's interpretation of this effect was that the final phoneme of the first word was being restored on the basis of lexical influences, and that the restored phoneme then triggered compensation for coarticulation, just as when the phoneme is unambiguous in the perceptual stimulus.

Advocates of bottom-up connectionist models have recently argued that, despite appearances, Elman and McClelland's (1988) results do not demonstrate top-down influence. Bottom-up connectionist models have been shown to be compatible with Elman and McClelland's results. For example, Norris (1993) presents results from a simulation involving a *simple recurrent network*, or SRN (introduced by Elman, 1988, 1990). As shown in Fig. 8.3, the SRN involves a crucial modification to a feedforward network: The current set of hidden unit values is "copied back" to

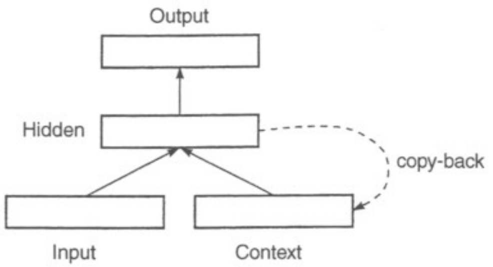

FIG. 8.3 Architecture of a simple recurrent network. This network is a conventional multi-layered feedforward network, except that there are additional input units into which the hidden unit activations from the previous time-step are copied.

a set of additional input units, and paired with the *next* input to the network. This means that the current hidden unit values can directly affect the next state of the hidden units; more generally, this means that there is a loop around which activation can reverberate over many time-steps. This gives the network a memory for past inputs, and therefore the ability to deal with integrated sequences of inputs presented successively. This contrasts with standard back-propagation networks, the behaviour of which is determined solely by the current input. This means that SRNs are able to tackle tasks such as language processing in which the input is revealed gradually over time, rather than being presented at once. For this reason SRNs have been widely used in connectionist models of language processing, as we shall see in subsequent sections, and there has also been some exploration of their computational properties (e.g. Chater & Conkey, 1992; Christiansen & Chater, in press; Cleeremans, Servan-Schreiber, & McClelland, 1989; Servan-Schreiber, Cleeremans, & McClelland, 1991).

Norris trained an SRN on input and output consisting of words (from a 12-word lexicon) presented one phoneme at a time. The input was represented in terms of phonetic features, which might have intermediate values, corresponding to ambiguous phonemes, and the output consisted of units, each detecting a particular phoneme. When input was presented to the trained network that had an ambiguous first word-final phoneme, and an ambiguous initial segment of the second word, a parallel to the compensation for coarticulation effect was observed, within the limits of the lexicon used: The percentages of /t/ and /k/ responses to the first phoneme of the second word depended on the identity of the first word, as in Elman and McClelland's original experiment. But the explanation for this pattern of results cannot be top-down influence from units representing words, since there are no units representing words in the network.

Norris's small-scale artificial example is no more than suggestive, however. The crucial question is: Would a network trained on natural speech, rather than on very small-scale artificial data, model Elman and McClelland's results? Shillcock et al. (1992) constructed such a network and found a close fit with Elman and McClelland's data. A recurrent network was trained on a corpus of phonologically transcribed conversational English, and inputs and outputs to the network were represented at the level of phonetic features. As in Norris's simulations, there is no lexical level of representation from which top-down information can flow. None the less, phoneme restoration follows the pattern that Elman and McClelland explain in terms of lexical influence.

Why is it that in the simulation purely bottom-up processes appear to mimic lexical effects? Shillcock et al. (1992) argue that restoration occurs in their network on the basis of statistical regularities at the phonemic level, rather than because of lexical influence. It just happens that the lexical items that Elman and McClelland used experimentally are more statistically regular at the phonemic level than the non-words with which they are contrasted. This is confirmed by a statistical analysis of the corpus of speech on which the network is trained. By carefully choosing stimulus items for which statistical regularities at the phonemic level have the opposite bias to that which would be provided by lexical status, it may be possible to experimentally distinguish between the interactive and bottom-up connectionist accounts. This experimental test is yet to be conducted, however.

The debate between interactive and bottom-up models of speech perception that we have just described is a good illustration of the way in which the introduction of connectionist models has led to unexpected theoretical predictions being derived (e.g. that bottom-up models can account for apparently lexically based phoneme restoration), as well as acting as a stimulus for empirical research.

Speech production

Throughout the field of language research as a whole, relatively little work has been done on the *production* of language. This general trend is also true of connectionist natural language processing. Thus, most connectionist language models address issues of processing and comprehension, rather than production. However, some steps have been taken towards the modelling of language production within a connectionist framework, most notably by Dell and colleagues (e.g. Dell, 1986; Dell, Juliano, & Govindjee, 1993; Dell & O'Seaghdha, 1991; Martin, Dell, Saffran, & Schwartz, 1994; Schwartz, Saffran, Bloch, & Dell, 1994).

Dell's (1986) spreading activation model of retrieval in sentence production constitutes one of the first connectionist attempts to account for speech production.[11] Although the model was presented as a sentence production model, only the phonological encoding of words was computationally implemented in terms of an interactive activation model. This lexical network consisted of hierarchically ordered layers of nodes, corresponding to the following linguistically motivated units: morphemes (or lexical nodes), syllables, rimes and consonant clusters, phonemes, and features. The individual nodes are connected bi-directionally to each other in a straightforward manner without lateral connections within layers, and with the exception of the addition of special null element nodes and syllabic position coding of nodes that correspond to syllables. For example, the lexical node for the word (morpheme) "spa" is connected to the /spa/ node in the syllable layer. The latter is linked to the consonant cluster /sp/ (onset) and the rime /a/ (nucleus). On the phoneme level, /sp/ is connected to /s/ (which in turn is linked to the features *fricative*, *alveolar*, and *voiceless*) and /p/ (which is connected to the features *bilabial*, *voiceless*, and *stop*). The rime /a/ is linked to the vowel /a/ in the phoneme layer (and subsequently is connected to the features *tense*, *low*, and *back*) and to a node signifying a null coda.

Processing begins with the activation of a lexical node (meant to correspond to the output from higher level morphological, syntactic, and semantic processing), and activation then gradually spreads downwards in the network. Activation also spreads upwards via the feedback connections. After a fixed period of time (determined by the speaking rate), the nodes with the highest activations are selected for the onset, vowel, and coda slots. Using this network model Dell was able to account for a variety of speech errors, such as substitutions (e.g. *dog* → *log*), deletions (*dog* → *og*), and additions (*dog* → *drog*). Speech errors occur in the model when an incorrect node becomes more active than the correct node (given the activated lexical node) and therefore gets selected instead. Such erroneous activation may be due to the feedback connections activating nodes other than those directly corresponding to the initial word node. Alternatively, other words in the sentence context as well as words activated as a product of internal noise may interfere with the processing of the network. This model also made a number of quantitative predictions concerning the retrieval of phonological forms during production, some of which were later confirmed experimentally in Dell (1988).

Dell's account of speech errors and the phonological encoding of words has had an impact on subsequent models of speech production, both the connectionist (e.g. Harley, 1993) as well as the more symbolic kind (e.g.

[11] A somewhat similar model of speech production was developed independently by Stemberger (1985). This model was inspired by the interactive activation framework of McClelland and Rumelhart (1981), whereas Dell's work was not.

Levelt, 1989). Nevertheless, Dell's model does suffer from a number of shortcomings, of which we mention a few here. As with the previously mentioned interactive activation models, the connections between the nodes on the various levels have to be hand-coded. This means that no learning is possible. In itself this is not a problem in principle if innate linguistic knowledge is assumed, but the information coded in Dell's model is language-specific and could not be innate. There is, however, a more urgent, practical side of this problem. It becomes very difficult to scale these models up, since every connection between each and every node has to be hand-coded. This shortcoming is alleviated by a recent recurrent network model presented by Dell et al. (1993). Their model learns to form mappings from lexical items to the appropriate sequences of phonological segments. The model consists of an SRN, as outlined earlier, with a small additional modification: The current *output*, as well as the current hidden unit state, was copied back as additional input to the network. This allowed both past activation states of the hidden unit layer as well as the output from the previous time-step to influence current processing.[12] When given an encoding of, for example, "can" as the lexical input the network was trained to produce the features of the first phonological segment /k/ on the output layer, then /æ/ followed by /n/, and then finally generate an end of word marker (null segment). Trained in this manner, Dell et al. were able to account for speech error data without having to build syllabic frames and phonological rules into the network (as was the case in Dell, 1986). Importantly, this recent connectionist model suggests that sequential biases and similarity may explain aspects of human phonology which have previously been attributed to separate phonological rules and frames. Furthermore, the model indicates that future speech production models may have to incorporate learning and distributed representations in order to accommodate the role that the entire vocabulary appears to play in phonological speech errors.[13] As with

[12] The idea of copying back output as part of the next input was first proposed by Jordan (1986).

[13] It should however be noted that despite the relative success of this feedforward learning model, Dell's (1986) interactive activation model is still a strong candidate as a model of speech production as it has a much wider empirical coverage. Moreover, it has been "lesioned" in 65 various ways to simulate language problems in aphasia. Schwartz et al. (1994) demonstrated that a reduction in connection strengths between nodes may provide a possible account of error patterns in jargon aphasia. Martin et al. (1994) showed that introducing an abnormally rapid decay rate for activated nodes allowed the model to simulate paraphasia in deep dysphasia, and that gradually changing this decay rate towards a normal level may simulate the pattern of recovery found in the longitudinal study of an aphasic patient. By changing connection strengths and/or the decay rate, Dell, Schwartz, Martin, Saffran, and Gagnon (1997) were in a similar manner able to fit the error patterns of 21 fluent aphasics as well as make a number of predictions that were subsequently confirmed.

reading, connectionist models of speech perception and production have modelled a wide range of empirical data on both normal and impaired performance. Moreover, they have contributed to a reassessment of fundamental theoretical issues and generated fresh experimental work on core theoretical questions, particularly on the question of whether speech perception and production are interactive or sequential. Connectionist research has greatly contributed to current thinking about speech perception and production. The validity of specific connectionist models, as well as the scope of connectionist approaches in general, will have an important role in shaping future research in this area.

MORPHOLOGICAL PROCESSING

One of the connectionist models that has created the most debate is Rumelhart and McClelland's (1986a) model of the learning of English past tense. This debate in many ways resembles the one following Seidenberg and McClelland's (1989) model of word recognition and naming discussed earlier. In particular, the debate has to a large extent focused on whether a single mechanism may be sufficient to account for the empirical data concerning the developmental patterns in English past-tense learning, or whether a dual-route mechanism is necessary. The discussion of the past-tense model also relates to the viability of the wickelfeature representation, which we have already described in the context of Seidenberg and McClelland's model of word naming. Here, we provide an overview of the current debate, as well as pointers to its wider ramifications.

Can a system without any explicit representation of rules account for rule-like behaviour? Rumelhart and McClelland's (1986a) model of the acquisition of the past tense in English was presented as an affirmative answer to this question. English past tense is an interesting test case because children very roughly appear to go through three stages during learning. In particular, children seemingly exhibit a pattern of U-shaped learning when acquiring English verbs and their past tenses. During the first stage, children only use a few verbs in past tense and these tend to be irregular words—such as *came*, *went*, and *took*—likely to occur with a very high frequency in the child's input. These verbs are furthermore mostly used in their correct past-tense form. At the second stage, children start using a much larger number of verbs in the past tense, most of these of the regular form, such as *pulled* and *walked*. Importantly, children now show evidence of rule-like behaviour. They are able to conjugate non-words, generating *jicked* as the past tense of *jick*, and they start to overgeneralise irregular verbs—even the ones they got right in stage 1—for example, producing *comed* or *camed* as the past tense of *come*. During the third stage, the children regain their ability to correctly form the past tense of irregular verbs while maintaining their correct conjugations of the regular verbs. Thus, it appears prima facie that children (as in the case of reading, discussed earlier) learn to use a rule-based route for dealing with regulars as well as non-words and a memorisation route for handling irregulars. But how can such seemingly dual-route behaviour be accommodated by a single mechanism employing just a single route?

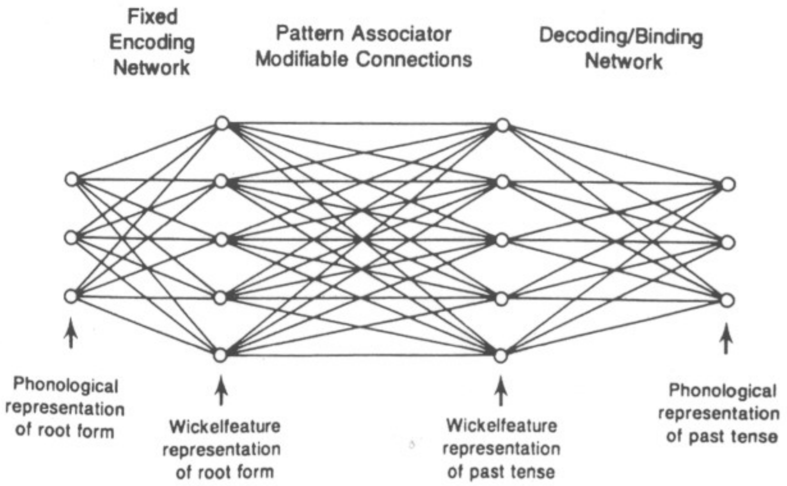
Fixed Encoding Network
Pattern Associator Modifiable Connections
Decoding/Binding Network
Phonological representation of root form
Wickelfeature representation of root form
Wickelfeature representation of past tense
Phonological representation of past tense

FIG. 8.4 The basic structure of the Rumelhart and McClelland (1986a) model for the learning of English past tense. Note that all learning takes place in the pattern associator. From "On learning the past tenses of English verbs" by D.E. Rumelhart and J.L. McClelland In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel Distributed Processing, Vol. 2* 1986b, MIT Press, p. 222. Copyright © (1986) MIT Press. Reprinted with permission.

Rumelhart and McClelland (1986a) showed that by varying the input to a connectionist model during learning, important aspects of the three stages of English past-tense acquisition could be simulated using a single mechanism. As illustrated in Fig. 8.4, the model consists of three parts: a fixed encoding network, a pattern associator network with modifiable connections, and a competitive decoding/binding network. The encoding network is an (unspecified) network, which takes phonological representations of root forms (presumably represented as wickelphones) and transforms them into a set of wickelfeatures. In order to promote generalisation, additional incorrect features are randomly activated, specifically, those features that have the same central feature as well as one of the two other context features in common with the input root form. The focus of interest in this model is the pattern associator network. It has 460 input and output units, each representing a wickelfeature. This network is trained to produce past-tense forms when presented with root forms of verbs as input. During training, the weights between the input and the output layers are modified using the perceptron learning rule (Rosenblatt, 1962) (the back-propagation rule is not required for this network, since it has just one modifiable layer). Since the output patterns of wickelfeatures generated by the association network most often do not correspond to a single past-tense form, the decoding/binding network must transform these distributed patterns into unique wickelphone representations. In this third network, each wickelphone in the 500 words used in the study was assigned to an output unit. These wickelphones compete individually for the input wickelfeatures in an iterative process. The more wickelfeatures a given wickelphone accounts for, the greater its strength. If two or more wickelphones account for the same wickelfeature the assigned "credit" is split between them in proportion to the number of other wickelfeatures they account for uniquely (i.e. a "rich get richer" competitive approach). The end result of this competition is a set of more or less non-overlapping wickelphones which correspond to as many as possible of the wickelfeatures in the input to the decoder network.

The pattern associator was trained in the following manner, showing evidence of going through the three relevant stages whilst learning English past tense. First, the net was trained on a set of 10 high-frequency verbs (8 irregular and 2 regular) for 10 epochs. At this point the net reached a satisfactory performance, treating both regular and irregular verbs in the same way (as also observed in the first stage of human acquisition of past tense). Next, 420 medium-frequency verbs (about 80% of these being regular) were added to the training set and the net was trained for an additional 190 epochs. Early on during this period of training the net behaved as children at acquisition stage 2: The net tended to regularise irregulars while getting regulars correct. At the end of the 190 epochs, network behaviour resembled that of children in stage 3 of the past-tense acquisition process, exhibiting an almost perfect performance on the 420 verbs. The network was then subsequently tested on a set of 86 low-frequency verbs (of which just over 80% were regular). The net appears to capture the basic U-shaped pattern of the acquisition of English past tense. In addition, it was able to exhibit differential performance on different types of irregular and regular verbs, effectively simulating some aspects of similar performance differences observed in children (Bybee & Slobin, 1982; Kuczaj, 1977, 1978). Moreover, the model demonstrated a reasonable degree of generalisation from the 420 verbs in the training set to the 86 low-frequency verbs in the test set; for example, demonstrating that it was able to use the three different regular endings correctly (i.e. using /t/ with root forms ending with an unvoiced consonant, /d/ as suffix to forms ending with a voiced consonant or vowel, and /ˆd/ with verb stems ending with a "t" or a "d").

The merits and inadequacies of the Rumelhart and McClelland (1986a) past-tense model have been the focus of much debate, originating with Pinker and Prince's (1988) detailed criticism (and to a lesser extent by Lachter & Bever's 1988 critique). Since then the debate has flourished across the symbolic/connectionist divide (e.g. on the symbolic side: Kim, Pinker, Prince, & Prasada, 1991; Pinker, 1991; and on the connectionist side: Cottrell & Plunkett, 1991; Daugherty, MacDonald, Petersen, & Seidenberg, 1993; MacWhinney & Leinbach, 1991; Seidenberg, 1992; Daugherty & Seidenberg, 1992). Here, we focus on the most influential aspects of the debate.

As was the case with the Seidenberg and McClelland (1989) model of reading, the use of wickelphones/wickelfeature representations has been considered problematic (e.g. by Pinker & Prince, 1988). Perhaps for this reason, most of the subsequent connectionist models of English past tense (both of acquisition, e.g. Plunkett & Marchman, 1991, 1993, and of diachronic change, Hare & Elman, 1992, 1995) therefore use a position-specific phonological representation in which vowels/consonants are defined in terms of phonological contrasts, such as voiced/unvoiced, front/centre/back. Another, more damaging, criticism of the single-route approach is that the U-shaped pattern of behaviour observed in the model during learning appears essentially to be an artifact of suddenly increasing the total number of verbs (from 10 to 420) in the second phase of learning. Pinker and Prince (1988) point out that no such sudden discontinuity appears to occur in the number of verbs to which children are exposed. Thus, the occurrence of U-shaped learning in the model is undermined by the psychological implausibility of the training regime.

More recently, however, Plunkett and Marchman (1991) showed that this training regime is not required to obtain U-shaped learning. They trained a feedforward network with a hidden unit layer on a vocabulary of artificial verb stems and past-tense forms, patterned by regularities patterned on the English past tense. They held the size of the vocabulary used in training constant at 500 verbs. They found that the

[14] In this connection, type frequency refers to the number of different words belonging to a given class, each counted once (e.g. the number of different regular verbs). Token frequency, on the other hand, denotes the number of instances of a particular word (e.g. number of occurrences of the verb *have*).

net not only was able to exhibit classical U-shaped learning, but also had learned various selective micro U-shaped developmental patterns observed in children's behaviour. For example, given a training set with a type and token frequency[14] reflecting that of English verbs the net was able to simulate a number of sub-regularities between the phonological form of a verb stem and its past-tense form (e.g. *sleep* → *slept*, *keep* → *kept*).[15] In a subsequent paper, Plunkett and Marchman (1993) obtained similar results using an incremental, and perhaps more psychologically plausible, training regime. Following initial training on 20 verbs, the vocabulary was gradually increased until reaching a size of 500 verb stems. This training regime significantly improved the performance of the net (compared with a similarly configured net trained on the same vocabulary in Plunkett and Marchman, 1991). This approach also suggested that a critical mass of verbs is needed before a change from rote-learning (memorisation) to system-building (rule-like generalisation behaviour) may occur—the latter perhaps related to the acceleration in the acquisition of vocabulary items (or "vocabulary spurt") observed when a child's overall vocabulary exceeds around 50 words (e.g. Bates, Bretherton, & Snyder, 1988).

Most recently, the connectionist models of past-tense acquisition have been accused of being too dependent on the token and type frequencies of irregular and regular vocabulary items in English. Prasada and Pinker (1993) have argued that the purported ability of connectionist models to simulate verb inflection may be an artifact of the idiosyncratic frequency statistics of English. The focus of the argument is the *default* inflection of words; for example, the *-ed* suffixation of English regular verbs. The default inflection of a word is assumed to be independent of its particular phonological shape and occurs unless the root form corresponds to a specific irregular form. According to Prasada and Pinker, connectionist models are dependent on frequency and surface similarity for their generalisation ability. In English, most verbs are regular, that is, many regular verbs have a high type frequency but a relatively low token frequency, allowing a network to construct a broadly defined default category. Irregular verbs in English, on the other hand, have a low type frequency but a high token frequency, the latter permitting the memorisation of the irregular past tenses in terms of a number of narrow phonological sub-categories (e.g. one for the *i–a* alternation in *sing* → *sang*, *ring* → *rang*, another for the *o–e* alternation in *grow* → *grew*, *blow* → *blew*, etc.). Prasada and Pinker (1993) show that the default generalisation in Rumelhart and McClelland's (1986a) model is dependent on a similar frequency distribution in the training set. They furthermore contend that no connectionist model can accommodate default generalisation for a class of words that has both low type frequency and low token frequency. The default inflection of plural nouns in German appear to fall in this category and would therefore seem to be outside the capabilities of connectionist networks (Clahsen, Rothweiler, Woest, & Marcus, 1993; Marcus, Brinkmann, Clahsen, Wiese, Woest & Pinker, 1993). If true, such lack of cross-linguistic validity would render neural network models of past tense acquisition obsolete.

[15] As pointed out by Pinker and Prince (1988), the Rumelhart and McClelland (1986a) model was not able adequately to accommodate such sub-regularities.

However, recent connectionist work has addressed this issue of minority default mappings with some success. Daugherty and Hare (1993) trained a feedforward network (with hidden units) to map the phonological representation of a stem to a phonological representation of the past tense given a set of verbs roughly representative of very early Old English (before about 870 AD). The training set consisted of five classes of irregular verbs plus one class of regular verbs—each class containing 25 words (each represented once in the training set). Thus, words taking the default generalisation /-ed/ formed a minority (i.e. only 17%) of the words in the training set. *Pace* Prasada and Pinker (1993) and others, the network was able to learn the appropriate default behaviour even when faced with a low-frequency default class. Indeed, it appears that generalisation in neural networks may not be strictly dependent on similarity to known items. Daugherty and Hare's (1993) results show that if the non-default (irregular) classes have a sufficient degree of internal structure, default generalisation may be promoted by the lack of similarity to known items. These results were corroborated by further simulations and analyses in Hare, Elman, and Daugherty (1995). Moreover, Forrester and Plunkett (1994) obtained similar results when training a feedforward model (with hidden units) to learn artificial input patterned on the Arabic plural. In Arabic, the majority of plural forms—called the Broken Plural—are characterised by a system of sub-regularities dependent on the phonological shape of the noun stem. In contrast, a minority of nouns takes the Sound Plural inflection which forms the default in Arabic. Forrester and Plunkett's net was trained to map phonological representations of the noun stems to their appropriate plural forms represented phonologically. Their results also indicate that connectionist models can learn default generalisation without relying on large word classes or direct similarity.

These positive results constitute important steps forward. Nevertheless, we presently have no detailed knowledge concerning the specific condition from which connectionist default generalisation can arise, nor do we know how it will scale when faced with the full complexity of language. On the other hand, rule-like and frequency-independent default generalisation may not be as pressing a problem for connectionist models as

Clahsen et al. (1993) and Marcus et al. (1993) claim. Via a reanalysis of the data concerning German noun inflection (in combination with additional data from Arabic and Hausa), Bybee (1995) showed that default generalisation is sensitive to type frequency and does not seem to be entirely rule-like. This kind of generalisation may fit better with the kind of default generalisation that connectionist models produce than with the rigid application of default rules in the symbolic models.

The issue of whether humans employ a single, connectionist-style mechanism for rule-like morphological processing is far from settled. Connectionist models can provide an impressive fit to a wide range of developmental and linguistic data. Even detractors of connectionist models of morphology typically allow that some kind of associative connectionist mechanism may explain the complex patterns found in the "irregular" cases. The controversial question is whether a single connectionist mechanism can simultaneously account both for regular and the irregular cases, or whether the regular cases can only be generated by a distinct route involving (perhaps necessarily symbolic) rules. The future is likely to bring further connectionist modelling of cross-linguistic data concerning morphology as well as a closer fitting of developmental micro patterns and distributional data to such models. As we shall see next, the question of whether language can be accounted for without the explicit representation of rules also plays an important part in connectionist modelling of syntactic processing.

SYNTAX

Syntactic processing is arguably the area of natural language which has the strongest ties to explicit rules as a means of explanation. Since Chomsky (1957), grammars have been understood predominately in terms of a set of generative phrase structure rules (often coupled with rules or principles for the further transformation of phrase structures). In early natural language research the central status of rules was directly reflected in the Derivational Theory of Complexity (Miller & Chomsky, 1963). This theory suggested that the application of a given rule (or transformation) could be measured directly in terms of time it takes for a listener/reader to process a sentence. This direct mapping between syntactic rules and response times was soon found to be incorrect, leading to more indirect ways of eliciting information about the use of rules in the processing of syntax. But can syntactic processing be accounted for without explicit rules? Radical connectionism aims to show that it can.

One way of dealing with syntax in connectionist models is to "hand-wire" symbolic structures directly into the architecture of the network. Much early work in connectionist processing of linguistic structure adopted this implementational approach; starting with Small, Cottrel, and Shastri's (1982) first attempt at connectionist parsing followed by Reilly's (1984) connectionist account of anaphor resolution and later by Fanty's (1985) connectionist context-free parser, Selman and Hirst's (1985) modelling of context-free parsing using simulated annealing, Waltz and Pollack's (1985) interactive model of parsing (and interpretation), McClelland and Kawamoto's (1986) neural network model of case-role assignment, and Miyata, Smolensky, and Legendre's (1993) structure-sensitive processing of syntactic structure using tensor representations (Smolensky, 1990). Such connectionist re-implementations of symbolic systems might have interesting computational properties and even be illuminating regarding the appropriateness of a particular style of symbolic model for distributed computation (Chater & Oaksford, 1990). On the other hand, there is the promise that connectionism may be able to do more than simply implement symbolic representations and processes; in particular, that networks may be able to *learn* to form and use structured representations. The most interesting models of this sort typically focus on learning quite limited aspects of natural language syntax. These models can be divided into two classes, depending on whether preprocessed sentence structures or simply bare sentences are presented as input.

The less radical class presupposes that the syntactic structure of each sentence to be learned is more or less given; that is, each input item is tagged with information pertaining the syntactic role of that item (e.g. the word *cat* may be tagged as Singular Noun). In this class we find, for example: connectionist parsers, such as PARSNIP (Hanson & Kegl, 1987) and VITAL (Howells, 1988); the structure dependent processing of Pollack's (1988, 1990) recursive auto-associative memory network subsequently used in Chalmers' (1990) model of active to passive transformation and in a model of syntactic processing in logic (Niklasson & van Gelder, 1994); Sopena's (1991) distributed connectionist parser incorporating attentional focus; and Stolcke's (1991) hybrid model deriving syntactic categories from phrase-bracketed examples given a vector space grammar. Typically, the task of these network models is to find the grammar (or part of thereof) which fits the example structures. This means that the structural aspects of language are not themselves learned by observation, but are built in. These models are related to statistical approaches to language learning such as stochastic context-free grammars (Brill, Magerman, Marcus, & Santorini, 1990; Jelinek, Lafferty, & Mercer, 1990) in which learning sets the probabilities of each grammar rule in a prespecified context-free grammar, from a corpus of parsed sentences.

The more radical models have taken on a much harder task, that of learning syntactic structure from strings of words, with no prior assumptions about the particular structure of the grammar. The most influential approach employs the earlier mentioned SRNs. It is fair to say that these radical models have so far reached only a modest level of performance. This may explain why the more radical connectionist attempts at syntax learning have not caused nearly as much debate as the earlier mentioned model of English past-tense acquisition (Rumelhart & McClelland, 1986a) and the model of reading aloud (Seidenberg & McClelland, 1989). Nevertheless, we focus on the radical connectionist models here because they potentially bear the promise of language learning without a priori built-in linguistic knowledge (*pace* e.g. Chomsky, 1965, 1986; Crain, 1991; Pinker, 1994; and many others).

Elman (1991, 1993) trained an SRN to predict the next word it will receive as input given sentences generated by a simple context-free grammar. This grammar involved subject noun/verb agreement, verbs with different argument structure (i.e. intransitive, transitive, and optionally transitive verbs), as well as subject and object relative clauses (allowing for multiple embeddings with complex long-distance dependencies). These simulations demonstrated that an SRN is able to acquire the grammatical regularities underlying a simple grammar. In addition, the SRN showed some behavioural similarities with human behaviour on centre-embedded structures (Weckerly & Elman, 1992). Christiansen (1994, in preparation) extended this work, training SRNs on more complex grammars involving prenominal genitives, prepositional modifications of noun phrases, noun phrase conjunctions, and sentential complements in addition to the grammatical features found in Elman's work. One of the grammars moreover incorporated cross-dependencies, a weakly context-sensitive structure found in languages such as Dutch and Swiss-German. Christiansen found that the SRNs were able to learn these more complex grammars, exhibiting the same kind of qualitative processing difficulties as humans do on similar sentence constructions (see also Christiansen & Chater, in press).

As we have seen, current models of syntax typically use "toy" fragments of grammar and small vocabularies. Aside from raising the question of the viability of scaling-up, this makes it difficult to provide detailed fits with empirical data. None the less, some attempts have more recently been made toward fitting existing data and deriving new empirical predictions from the models. For example, Tabor, Juliano, and Tanenhaus (1997) present a SRN-based dynamic parsing model that fits reading time data concerning the interaction between lexical and structural constraints in the resolution of temporary syntactic ambiguities (i.e. garden-path effects) in sentence comprehension. MacDonald and Christiansen (submitted) provide SRN simulations of reading time data concerning the differential processing of singly centre-embedded subject and object relative clauses by good and poor comprehenders. Finally, Christiansen (in preparation; Christiansen & Chater, in press) describes an SRN trained on recursive sentence structures, which fits grammaticality ratings data from several behavioural experiments. He also derives novel predictions about the processing of sentences involving multiple prenominal genitives, multiple prepositional modifications of nouns, and doubly centre-embedded object relative clauses, which have subsequently been empirically confirmed (Christiansen & MacDonald, in preparation).

These simulation results suggest that SRNs may be viable models of syntactic processing. However, connectionist models of language learning (i.e. Chalmers, 1990; Elman, 1990; McClelland & Kawamoto, 1986; Miyata et al. 1993; Pollack, 1990; Smolensky, 1990; St. John & McClelland, 1990) have recently been attacked for not affording the kind of generalisation abilities that would be expected from models of language. Hadley (1994a) correctly pointed out that generalisation in much connectionist research has not been viewed in a sophisticated fashion. The testing of generalisation is typically done by recording network output given a test set consisting of items not occurring in the original training set, but potentially containing many similar structures and word sequences. Hadley insisted that to demonstrate genuine, "strong" generalisation a network must be shown to learn a word in one syntactic position and then generalise to using/processing that word in another, novel syntactic position. He challenged connectionists to adopt a more rigorous training and testing regime in assessing whether networks really generalise successfully in learning syntactically structured material.

Christiansen and Chater (1994) addressed this challenge, providing a formalisation of Hadley's original ideas as well as presenting evidence that connectionist models are able to attain strong generalisation. In their training corpus (generated by the grammar from Christiansen, 1994), the noun *boy* was prevented from ever occurring in a noun phrase conjunction (i.e. noun phrases such as *John and boy* and *boy and John* did not occur). During training the SRN had therefore only seen singular verbs following *boy*. None the less, the net was able to predict correctly that a plural verb must follow *John and boy* as prescribed by the grammar. In addition, the net was still able to predict correctly a plural verb when a prepositional phrase was attached to *boy* as in *John and boy from town*, providing even stronger evidence for strong generalisation. This suggests that the SRN is able to make non-local generalisations based on the structural regularities in the training corpus (see Christiansen & Chater, 1994, for further details). If the SRN relied solely on local information it would not have been able to make correct predictions in either case. Christiansen (in preparation) demonstrated that the same SRN also was able to generalise appropriately when presented with completely novel words, such as *zorg*,[16] in a noun phrase conjunction by predominately activating the plural verbs. In contrast, when the SRN was presented with ungrammatical lexical items in the second noun position, as in *John and near*, it did not activate the plural nouns. Instead, it activated lexical items that were not grammatical given the previous context. The SRN was able to generalise to the use of known words in novel syntactic positions as well as to the use of completely novel words. At the same time, it was also able to distinguish items that were grammatical given previous context from those that were not. Thus, the network demonstrated sophisticated generalisation abilities, ignoring local word co-occurrence constraints, while appearing to comply with structural information at the constituent level. Additional evidence of strong generalisation in connectionist nets are found in Niklasson and van Gelder (1994; but see Hadley, 1994b for a rebuttal).

One possible objection to these models of syntax is that connectionist (and other bottom-up statistical) models of language learning will not be able to scale up to solve human language acquisition because of arguments pertaining to the purported poverty of the stimulus (see Seidenberg, 1994 for a discussion). However, there is evidence that some models employing simple statistical analysis may be able to scale up and even attain strong generalisation. When Redington, Chater, and Finch (1993) applied a method of distributional statistics (see also Finch & Chater, 1992, 1993) to a corpus of child-directed speech (the CHILDES corpus collected by MacWhinney & Snow, 1985), they found that the syntactic category of a new word could be derived from a single occurrence of that word in the training corpus. This indicates that strong generalisation may be learnable through the kinds of bottom-up statistical analysis that connectionist models appear to employ—even on a scale comparable with that of a child learning her first language. In this context, it is also important to note that achieving strong generalisation is not only a problem for learning-based connectionist models of syntactic processing. As pointed out by Christiansen and Chater (1994), most symbolic models cannot be ascribed strong generalisation since in most cases they are spoon-fed the lexical categories of words via syntactic tagging. The question of strong generalisation is therefore just as pressing for symbolic approaches as for connectionist approaches to language acquisition. The results outlined here suggest that connectionist models may be closer to solving this problem than their symbolic counterparts.

[16] In these simulations novel words corresponded to units that had not been activated during training.

Other aspects of language processing

There are a number of areas within connectionist natural language processing that have not received attention in this chapter. Amongst these are, for instance, models that deal with semantic aspects of language, models that address various issues at the level of discourse, and hybrid models seeking to combine the best of both the connectionist and the symbolic world. Unfortunately, space does not allow us to discuss such models here. Instead, we provide a few pointers for further reading.

Various aspects of semantic processing have been addressed in connectionist models, such as: word sense disambiguation (Cottrell, 1985); disambiguation of prepositional-phrase attachments using soft lexical preference rules (Sharkey, 1992); and incremental interpretation via the learning and application of contextual constraints in sentence comprehension (St. John & McClelland, 1990). Connectionist models of discourse and text comprehension include, for example, Allen's (1990) use of modified SRNs (called "connectionist language users") in a simple question answering task; Karen's (1990) modified SRN model of topic identification from written narrative discourse; and Sharkey's (1990) model of text comprehension involving four network modules (for respectively goals/plans, sequencing, knowledge, and the lexicon). Recent years have seen a surge in the number of hybrid connectionist/symbolic models of which we mention but a few examples: Bourlard and Morgan (1994) employ multi-layered feedforward networks to boost the performance of a state-of-the-art automatic speech recognition system based on hidden Markov models; Kwasny and Faisal (1990) implement a deterministic Marcus (1980) style parser in which a feedforward network is trained to suggest parsing actions given the state of a symbolic stack and buffer (see also Kwasny, Johnson, & Kalman, 1994, in which the feedforward net is replaced with an SRN); and Miikkulainen (1993) assembles modular networks dedicated to aspects of lexical, sentence, and story processing in a model of text comprehension inspired by symbolic script-based systems.

Overall, connectionist models of syntax and higher level aspects of language processing remain in early stages of development, and have not attained the level of sophistication of connectionist accounts of speech perception, production, reading, or morphology. Future research is required to decide whether promising, but limited, initial results can eventually be scaled up to deal with the complexities of real language input, or whether a purely connectionist approach is beset by fundamental limitations, and can only succeed to the extent that it rediscovers and reimplements the symbolic representations postulated by generative linguistics.

CONCLUSION

We have seen that controversy surrounds both the current significance of, and future prospects for, connectionist models of language processing. Current connectionist models involve over-simplifications with respect to the full complexity of human natural language processing, and only future research will determine the extent to which current models can be "scaled-up" successfully. Connectionism has, however, already influenced theoretical debates within the psychology of language processing in a number of ways, and we outline some of these influences here.

First, connectionist models have provided the first fully explicit and psychologically relevant computational models in a number of language processing domains, such as reading and past tense learning. Previous accounts in these areas consisted of "box-and-arrow" flow diagrams rather than detailed computational mechanisms. Whatever the lasting value of connectionist models themselves, they have certainly raised the level of theoretical debate in these areas, by challenging theorists of all viewpoints to provide computationally explicit accounts.

Second, the centrality of learning in connectionist models has brought a renewed interest in mechanisms of language learning (Bates & Elman, 1993), while Chomsky (e.g. 1986) has argued that although there are "universal" aspects of language that are innate, the vast amount of information specific to the language that the child acquires must be learned. Connectionist models provide mechanisms for how (at least some of) this learning might occur, whereas previous symbolic accounts of language processing have not taken account of how learning might occur. Furthermore, the attempt to use connectionist models to learn syntactic structure encroaches on the area of language for which Chomsky has argued innate information must be central. The successes and failures of this programme thus directly bear on the validity of this viewpoint.

Third, the dependence of connectionist models on statistical properties of their input has been one contributory factor in the upsurge of interest in the role of statistical factors in language learning (MacWhinney, Leinbach, Taraban, & McDonald, 1989; Redington et al. 1993) and language processing. This renewed interest in statistics is, of course, entirely compatible with the view that language processing takes account of structural properties of language, as described by classical linguistics. More radical connectionists have, as we have noted, also attempted to encroach on the territory of classical linguistics.

Finally, connectionist systems have given rise to renewed theoretical debate concerning what it really means for a computational mechanism to implement a rule, whether there is a distinction between "implicit" and "explicit" rules (see e.g. Davies, 1995 for discussions), and which kind should be ascribed to the human language processing system.

Connectionism has, we suggest, already had an important influence on the development of the psychology of language. But the final extent of that influence depends on the degree to which practical connectionist models can be developed and extended to deal with complex aspects of language processing in a psychologically realistic way. If realistic connectionist models of language processing can be provided, then the possibility of a radical rethinking not just of the nature of language processing, but of the structure of language itself, may be required. It might be that the ultimate description of language resides in the structure of complex networks, and can only be approximately expressed in terms of structural rules, in the style of generative grammar. On the other hand, it may be that connectionist models can only succeed to the extent that they build in standard linguistic constructs, or that connectionist learning methods do not scale up at all. The future development of connectionist models of language is therefore likely to have important implications for the theory of language processing and language structure, either in overturning, or reaffirming, traditional psychological and linguistic assumptions.

FURTHER READING

The suggested readings are grouped according to the general structure of the chapter.

Background. The PDP volumes (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986b) provide a solid introduction to application of neural networks in cognitive models. Smolensky (1988) offers a connectionist alternative to viewing cognition as symbol manipulation, whereas Fodor and Pylyshyn (1988) is a classic criticism of connectionism.

Visual word recognition and word naming. For an early interactive activation model of visual word recognition, see McClelland & Rumelhart (1981). Seidenberg and McClelland (1989) is a classic paper on connectionist models of reading. Coltheart et al. (1993) provide a criticism of this model and a symbolic alternative. For the most recent advancement of this discussion, see Plaut et al. (1996).

Lexical processing during speech. The TRACE model of speech perception is described in McClelland and Elman (1986). The classic model of speech production and speech errors is Dell (1986).

Morphological processing. Rumelhart and McClelland (1986a) and Pinker & Prince (1988) define the two sides of the past-tense debate. See Plunkett and Marchman (1993) and Pinker (1991) for recent updates.

Syntax. Elman (1993) provides a recent update on an influential connectionist approach to the learning of syntactic regularities, but see Hadley (1994a) for a criticism of this and other connectionist models of syntax. For a survey of the most recent research on connectionist language processing—including discussions of its future prospects—see Christiansen et al. (in press).

ACKNOWLEDGEMENTS

We would like to thank Gary Dell, Joe Devlin, Kim Plunkett, and Ronan Reilly for commenting on an earlier draft of this chapter. NC is partially supported by the Joint Councils Initiative in Cognitive Science/HCI, Grant SPG 9029590. Part of the research presented here was carried out while MHC was a McDonnell Postdoctoral Fellow in the PNP Program at Washington University in St. Louis.

REFERENCES

Aderman, D., & Smith, E.E. (1971). Expectancy as a determinant of functional units in perceptual cognition. *Cognitive Psychology*, *2*, 117–129.

Allen, R.B. (1990). Connectionist language users. *Connection Science*, *2*, 279–311.

Ashby, W.R. (1952). *Design for a brain.* New York: John Wiley & Sons.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first word to grammar: Individual differences and dissociable mechanisms.* Cambridge: Cambridge University Press.

Bates, E.A., & Elman, J.L. (1993). Connectionism and the study of change. In M.J. Johnson (Ed.), *Brain development and cognition* (pp. 623–642). Cambridge, MA: Basil Blackwell.

Besner, D., Twilly, L., McCann, R.S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review*, *97*, 432–446.

Boole, G. (1854). *The laws of thought.* London: Macmillan.

Bourlard, H.A., & Morgan, N. (1994). *Connectionist speech recognition: A hybrid approach.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brill, E., Magerman, D., Marcus, M., & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. In *DARPA speech and natural language workshop.* Hidden Valley, PA: Morgan Kaufmann.

Bruner, J. (1957). On perceptual readiness. *Psychological Review*, *64*, 123–152.

Bryson, A.E., & Ho, Y.C. (1975). *Applied optimal control.* New York: Hemisphere.

Bullinaria, J.A. (1994). *Representation, learning, generalisation and damage in neural network models of reading aloud* (Tech. Rep.). Edinburgh, UK: University of Edinburgh, Department of Psychology.

Bullinaria, J.A., & Chater, N. (1995). Connectionist modelling: Implications for neuropsychology. *Language and Cognitive Processes*, *10*, 227–264.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

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

Chalmers, D.J. (1990). Syntactic transformations on distributed representations. *Connection Science*, *2*, 53–62.

Chater, N., & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 402–407). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Chater, N., & Oaksford, M. (1990). Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition*, *34*, 93–107.

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

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Knowledge of language.* New York: Praeger.

Christiansen, M. (1994). *Infinite languages, finite minds: Connectionism, learning and linguistic structure.* Unpublished doctoral dissertation, University of Edinburgh, UK.

Christiansen, M.H. (in preparation). *Intrinsic constraints on the processing of recursive sentence structure.*

Christiansen, M.H., & Chater, N. (1992). Connectionism, meaning and learning. *Connection Science*, *4*, 227–252.

Christiansen, M., & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, *9*, 273–287.

Christiansen, M.H.M, & Chater, N. (in press). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science.*

Christiansen, M.H., Chater, N., & Seidenberg, M.S. (Eds.). (in press). Connectionist models of language processing: Progress and prospects. *Cognitive Science.*

Christiansen, M.H., & MacDonald, M.C. (in preparation). *Processing of recursive sentence structure: Testing predictions from a connectionist model.*

Churchland, P.S., & Sejnowski, T.J. (1989). Neural representation and neural computation. In L. Nadel, L. Cooper, P. Culicover & R.M. Harnish (Eds.), *Neural connections, mental computations.* Cambridge, MA: MIT Press.

Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G.F. (1993). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, *45*, 225–255.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

Cole, R.A., & Jakimik, J. (1978). Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of information processing.* New York: Academic Press.

Cole, R.A., & Jakimik, J. (1980). A model of speech perception. In R.A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–164). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Coltheart, M. (1985). Cognitive neuropsychology and the study of reading. In M.I. Posner & O.S.M. Marin (Eds.), *Attention and performance, XI. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.*

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Cottrell, G.W. (1985). *A connectionist approach to word sense disambiguation* (Tech. Rep. No. TR154). Rochester, NY: University of Rochester, Department of Computer Science.

Cottrell, G.W., & Plunkett, K. (1991). Learning the past tense in a recurrent network: Acquiring the mapping from meanings to sounds. In *Proceedings of the 13th annual meeting of the Cognitive Science Society* (pp. 328–333). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 597–650.

Daugherty, K., & Hare, M. (1993). What's in a rule? The past tense by some other name might be called a connectionist net. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman,

& A. Weigand (Eds.), *Proceedings of the 1993 connectionist models summer school* (pp. 149–156). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Daugherty, K., MacDonald, M., Petersen, A.S., & Seidenberg, M.S. (1993). Why no mere mortal has ever flown out to center field, but often people say they do. In *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 383–388). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Daugherty, K., & Seidenberg, M.S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 259–264). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Davies, M. (1995). Two notions of implicit rules. In J.E. Tomberlin (Ed.), *Philosophical perspectives: Vol. 9: AI, connectionism, and philosophical psychology*. Atascadero, CA: Ridgeview Publishing Company.

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

Dell, G.S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language, 27*, 124–142.

Dell, G.S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149–195.

Dell, G.S., & O'Seaghdha, P.G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1990). *Psychological Review, 98*, 604–614.

Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., & Gagnon, D.A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*, 801–838.

Elman, J.L. (1988). *Finding structure in time* (Tech. Rep. No. CRL-8801). San Diego, CA: University of California, Center for Research in Language.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Elman, J.L. (1991). Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Elman, J.L., & McClelland, J.L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143–165.

Fanty, M. (1985). *Context-free parsing in connectionist networks* (Tech. Rep. No. TR-174). Rochester, NY: University of Rochester, Department of Computer Science.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories by unsupervised learning. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 820–825). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Finch, S., & Chater, N. (1993). Learning syntactic categories: A statistical approach. In M. Oaksford & G.D.A. Brown (Eds.), *Neurodynamics and psychology* (pp. 295–321). New York: Academic Press.

Fodor, J.A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

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

Forrester, N., & Plunkett, K. (1994). Learning the Arabic plural: The case for minority mappings in connectionist networks. In *Proceedings of the 16th annual meeting of the Cognitive Science Society* (pp. 319–324). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Forster, K.I., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627–635.

Fox, R.A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 526–540.

Funnell, E. (1983). Phonological processing in reading: New evidence from acquired dyslexia. *British Journal of Psychology, 74*, 159–180.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–115.

Hadley, R.F. (1994a). Systematicity in connectionist language learning. *Mind and Language, 9*, 247–272.

Hadley, R.F. (1994b). Systematicity revisited: Reply to Christiansen & Chater and Niklasson & van Gelder. *Mind and Language, 9*, 431–444.

Hanson, S.J., & Kegl, J. (1987). PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. In *Proceedings of the 8th annual meeting of the Cognitive Science Society* (pp. 106–119). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hare, M., & Elman, J.L. (1992). A connectionist account of English inflectional morphology: Evidence from language change. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 265–270). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hare, M., & Elman, J.L. (1995). Learning and morphological change. *Cognition, 56*, 61–98.

Hare, M., Elman, J.L., & Daugherty, K.M. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes, 10*, 601–630.

Harley, T.A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes, 8*, 291–309.

Harm, M., Altmann, L., & Seidenberg, M. (1994). Using connectionist networks to examine the role of prior constraints in human learning. In *Proceedings of the 16th annual conference of the Cognitive Science Society* (pp. 392–396). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Hinton, G.E., & Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing*, Vol. 1. (pp. 282–317). Cambridge, MA: MIT Press.

Howells, T. (1988). VITAL, a connectionist parser. In *Proceedings of the 10th international conference on computational linguistics*, Stanford, CA.

Jared, D., McRae, K., & Seidenberg, M.S, (1990). The basis of consistency effect effects in word naming. *Journal of Memory and Language, 29*, 687–715.

Jelinek, F., Lafferty, J.D., & Mercer, R.L. (1990). *Basic methods of probabilistic context free grammars* (Tech. Rep. No. RC 16374 72684). Yorktown Heights, NY: IBM.

Johnston, J.C., & McClelland, J.L. (1973). Visual factors in word perception. *Perception and Psychophysics, 14*, 365–370.

Jordan, M. (1986). *Serial order: A parallel distributed approach* (Tech. Rep. No. 8604). San Diego, CA: University of California, San Diego, Institute for Cognitive Science.

Karen, L.F.R. (1990). Identification of topical entities in discourse: A connectionist approach to attentional mechanisms in language. *Connection Science, 2*, 103–122.

Kim, J.J., Pinker, S., Prince, S., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science, 15*, 173–218.

Koch, C., & Segev, I. (Eds.) (1989). *Methods in neuronal modeling: From synapses to networks*. Cambridge, MA: MIT Press.

Kucera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Kuczaj, S.A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16*, 589–600.

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

Kwasny, S.C., & Faisal, K.A. (1990). Connectionism and determinism in a syntactic parser. *Connection Science, 2*, 63–82.

Kwasny, S.C., Johnson, S., & Kalman, B.L. (1994). Recurrent natural language parsing. In *Proceedings of the 16th annual meeting of the Cognitive Science Society* (pp. 525–530). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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

Levelt, W.J.M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

MacDonald, M.C., & Christiansen, M.H. (submitted). *Individual differences without working memory: A reply to Just & Carpenter and Waters & Caplan.*

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

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255–277.

MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language, 12*, 271–295.

Mann, V.A. & Repp, B.H. (1980). Influence of vocalic context of perception of the [s]–[d] distinction. *Perception and Psychophysics, 28*, 213–228.

Marcus, M. (1980). *A theory of syntactic recognition for natural language.* Cambridge, MA: MIT Press.

Marcus, G.F., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1993). *German inflection: The exception that proves the rule* (MIT Occasional Paper No. 47). Cambridge, MA: MIT, Department of Brain and Cognitive Sciences.

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

Marshall, J.C., & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research, 2*, 175–199.

Marslen-Wilson, W.D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–523.

Marslen-Wilson, W.D., & Tyler, L.K. (1975). Processing structure of sentence perception. *Nature, 257*, 784–786.

Martin, N., Dell, G.S., Saffran, E.M., & Schwartz, M.F. (1994). Origins of paraphasia in deep dysphasia: Testing the consequence of decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language, 47*, 609–660.

Massaro, D.W. (1981). Sound to representation: An information-processing analysis. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 181–193). New York: North-Holland.

McClelland, J.L., & Elman, J.L. (1986). Interactive processes in speech perception: The TRACE model. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 2* (pp. 58–121.) Cambridge, MA: MIT Press.

McClelland, J.L., & Johnston, J.C. (1977). The role of familiar units in the perception of words and non-words. *Perception and Psychophysics, 22*, 249–261.

McClelland, J.L., & Kawamoto, A.H. (1986). Mechanisms of sentence processing. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 2* (pp. 272–325). Cambridge, MA: MIT Press.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Pt. 1. An account of basic findings. *Psychological Review, 88*, 375–407.

McClelland, J.L., & Rumelhart, D.E. (Eds.). (1986). *Parallel distributed processing: Vol. 2. Psychological and biological models.* Cambridge, MA: MIT Press.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon and memory.* Cambridge, MA: MIT Press.

Miller, G.A., & Chomsky, N. (1963). Finitary models of language users. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Vol. II* (pp. 419–491). New York: John Wiley & Sons .

Minsky, M. (1954). *Neural nets and the brain-model problem.* Unpublished doctoral dissertation, Princeton University, NJ.

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

Miyata, Y., Smolensky, P., & Legendre, G. (1993). Distributed representation and parallel distributed processing of recursive structures. In *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 759–764). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Morton, J., & Patterson, K.E. (1980). A new attempt at an interpretation, or, an attempt at a new interpretation. In M. Coltheart, K.E. Patterson & J.C. Marshall (Eds.), *Deep dyslexia.* London: Routledge.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Niklasson, L., & van Gelder, T. (1994). On being systematically connectionist. *Mind and Language, 9*, 288–302.

Norris, D.G. (1993). Bottom-up connectionist models of "interaction". In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Patterson, K.E., Seidenberg, M.S., & McClelland, J.L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R.G.M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). Oxford: Oxford University Press.

Pearlmutter, B.A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1*, 263–269.

Pinker, S. (1991). Rules of language. *Science, 253*, 530–535.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: William Morrow & Company.

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

Pisoni, D.B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics, 15*, 285–290.

Plaut, D., & McClelland, J.L. (1993). Generalization with componential attractors: Word and non-word reading in an attractor network. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Plaut, D., McClelland, J.L., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Plunkett, K. (1995). Connectionist approaches to language acquisition. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 36–72). Cambridge, MA: Basil Blackwell.

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

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition, 48*, 21–69.

Pollack, J.B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the 10th annual meeting of the Cognitive Science Society* (pp. 33–39). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Pollack, J.B. (1990). Recursive distributed representations. *Artificial Intelligence, 46,* 77–105.

Prasada, S., & Pinker, S. (1993). Similarity-based and rule-based generalizations in inflectional morphology. *Language and Cognitive Processes, 8,* 1–56.

Redington, M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 848–853). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Reilly, R.G. (1984). A connectionist model of some aspects of anaphor resolution. In *Proceedings of the 10th international conference on Computational Linguistics,* Stanford, CA.

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

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). *Learning internal representations by error propagation.* In J.L. McClelland. & D.E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 1* (pp. 318–362). Cambridge, MA: MIT Press.

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

Rumelhart, D.E., & McClelland, J.L. (1986a). On learning of past tenses of English verbs. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 2* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (Eds.) (1986b). *Parallel distributed processing: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 195–248). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Schwartz, M.F., Saffran, E.M., Bloch, D., & Dell, G.S. (1994). Disordered speech production in aphasic and normal speakers. *Brain and Language, 47,* 52–88.

Seidenberg, M.S. (1992). Connectionism without tears. In S. Davis (Ed.), *Connectionism: Advances in theory and practice* (pp. 84–122). Oxford, UK: Oxford University Press.

Seidenberg, M.S. (1994). Language and connectionism: The developing interface. *Cognition, 50,* 385–401.

Seidenberg, M.S., & Harm, M. (1995, November). *Division of labor and masking in a multicomponent model of word recognition.* Paper presented at the 36th annual meeting of the Psychonomics Society, Los Angeles.

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

Seidenberg, M.S., & McClelland, J.L. (1990). More words but still no lexicon: Reply to Besner et al. (1990). *Psychological Review, 97,* 447–452.

Seidenberg, M.S., Waters, G.S., Barnes, M.A., & Tanenhaus, M.K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23,* 383–404.

Sejnowski, T.J. (1986). Open questions about computation in the cerebral cortex. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing, Vol. 2* (pp. 372–389). Cambridge, MA: MIT Press.

Sejnowski, T.J., & Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145–168.

Selman, B., & Hirst, G. (1985). A rule-based connectionist parsing system. In *Proceedings of the 7th annual meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7,* 161–193.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge: Cambridge University Press.

Sharkey, N.E. (1990). A connectionist model of text comprehension. In D.A. Balota, G.B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Sharkey, N.E. (1991). Connectionist representation techniques. *AI Review, 5,* 143–167.

Sharkey, N.E. (1992). Functional compositionality and soft preference rules. In B. Linggard & C. Nightingale (Eds.), *Neural networks for images, speech, and natural language.* London: Chapman & Hall.

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 408–413). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Small, S.L., Cottrell, G.W., & Shastri, L. (1982). Towards connectionist parsing. In *Proceedings of the national conference on Artificial Intelligence.* Pittsburgh, PA.

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

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46,* 159–216.

Sopena, J.M. (1991). *ERSP: A distributed connectionist parser that uses embedded sequences to represent structure* (Tech. Rep. No. UB-PB-1-91). Barcelona, Spain: Universitat de Barcelona, Departament de Psicologia Bàsica.

Stemberger, J.P. (1985). An interactive activation model of language production. In A.W. Ellis (Ed.), *Progress in the psychology of language, Vol. 1* (pp. 143–186). Hove, UK: Lawrence Erlbaum Associates Ltd.

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

Stolcke, A. (1991). Syntactic category formation with vector space grammars. In *Proceedings from the 13th annual conference of the Cognitive Science Society* (pp. 908–912). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Tabor, W., Juliano, C., & Tanenhaus, M.K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12,* 211–271.

Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word recognition. *Journal of Memory and Language, 26,* 608–631.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2, 42,* 230–265.

Waltz, D.L., & Pollack, J.B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9,* 51–74.

Weckerly, J., & Elman, J. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 414–419). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Werbos, P.J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Unpublished doctoral dissertation. Cambridge, MA: Harvard University.

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