2

Connectionist Psycholinguistics in Perspective

Morten H. Christiansen and Nick Chater

Connectionist approaches to language have been, and still are, highly controversial. 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; it also 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 behavior 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 this chapter we aim to set the scene for the present volume on connectionist psycholinguistics, providing a brief historical and theoretical background as well as an update on current research in the specific topic areas outlined later. First we describe the historical and intellectual roots of connectionism, then introduce the elements of modern connectionism and 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 five central topics within connectionist psycholinguistics: speech processing, morphology, sentence processing, language production, and reading. We evaluate the research in each of these areas in terms of the three criteria for connectionist psycholinguistics discussed in Chapter 1: data contact, task veridicality, and input representativeness. The five topics illustrate the range of connectionist research on language discussed in more depth in the other chapters in Part I of this volume. They also provide an opportunity to assess the strengths and weaknesses of connectionist methods across this range, setting the stage for the general debate concerning the validity of connectionist methods in Part II of this volume. Finally, we sum up and consider the prospects for future connectionist research, and its relation to other approaches to the understanding of language processing and linguistic structure.

BACKGROUND

From the perspective of modern cognitive science, we tend to see theories of human information processing as borrowing from theories of information processing in machines (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 intellectual tradition of viewing human thought as 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 modeled by symbolic operations on a tape (the Turing machine), through von Neumann's motivation for the 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.

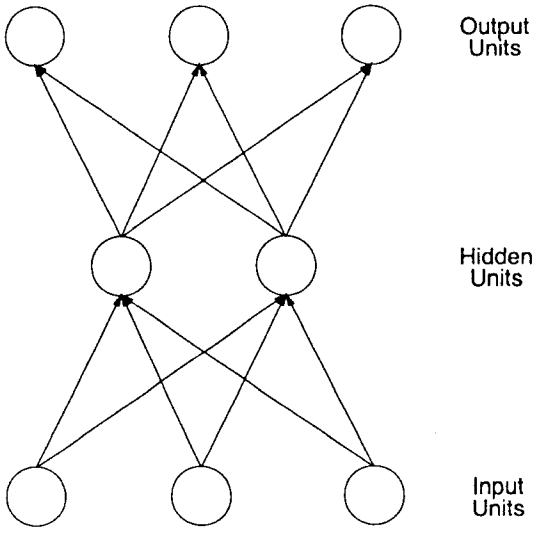
Connectionism (also known as "parallel distributed processing," "neural networks," or "neurocomputing") can be traced to a different tradition,

which attempts to design computers inspired by the structure of the brain.[1] McCulloch and Pitts (1943) provided an early and influential idealization of neural function. In the 1950s and 1960s Ashby (1952), Minsky (1954), Rosenblatt (1962), and others designed computational schemes based on related idealizations. Aside from their biological origin, these schemes were of interest because they were able to learn from experience, rather than being designed. Such "self-organizing" or learning machines therefore seemed plausible as models of learned cognitive abilities, including many aspects of language processing (although Chomsky, 1965, among others, challenged the extent to which language is learned). Throughout this period connectionist and symbolic computation stood as alternative paradigms for modeling intelligence, and it was unclear which would prove to be the most successful. But gradually the symbolic paradigm gained ground, providing powerful models in core domains such as language (Chomsky, 1965) and problem solving (Newell & Simon, 1972). Connectionism was largely abandoned, particularly in view of the limited power of then current connectionist methods (Minsky & Papert, 1969). But more recently, some of these limitations have been overcome (e.g., Hinton & Sejnowski, 1986; Rumelhart, Hinton, & Williams, 1986), reopening the possibility that connectionism 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. These neurons do not appear to tackle information processing problems alone. Rather, large numbers of neurons operate cooperatively 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 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 cooperatively 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.

The most popular of the connectionist networks is the *feed-forward network*, as illustrated in Figure 2.1. In this type of network the units are divided into "layers" and activation flows 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 (HU). The activation of each unit is determined by its current input (calculated as

Figure 2.1
Feed-Forward Network



Information flows entirely bottom-up in these networks, from the input units through the hidden units to the output units, as indicated by the arrows.

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 feed-forward connectionist networks.

Feed-forward networks learn from exposure to examples, and learning is typically achieved using the back-propagation learning algorithm (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 the overall level of error, averaged across input–target pairs. Back-propagation is a procedure that 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 small, meaning that the network produces the desired output. Notice that the network will produce an output

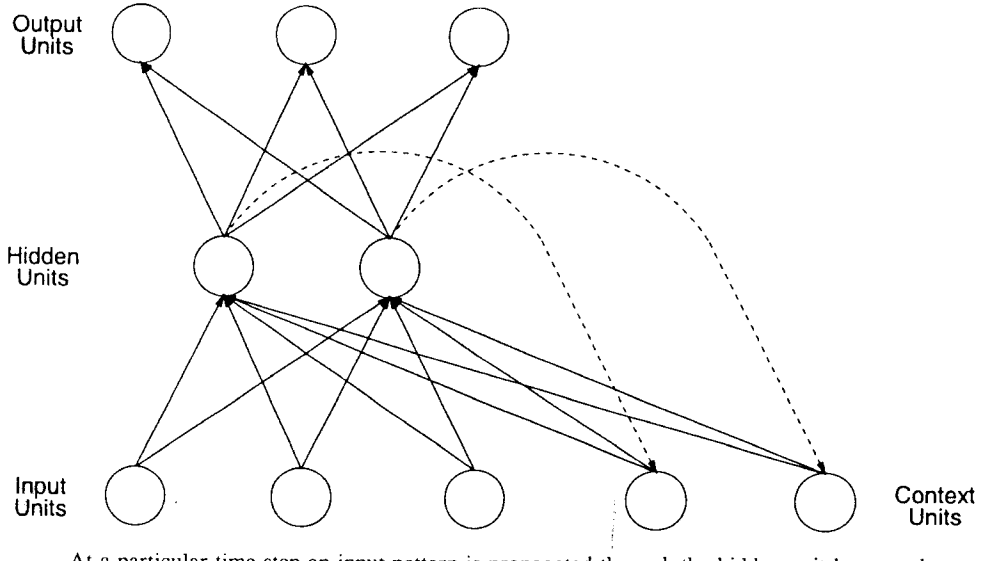
not only for inputs on which it has been trained, but for any input. If the network has learned about regularities in the mapping between inputs and targets, then it should be able to generalize successfully (i.e., to produce appropriate outputs in response to these new inputs).

Back-propagation may sound too good to be true. 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, even though a quite different configuration of weights would give rise to much lower error, if only it could be found by the learning process. The network is stuck in a *local minimum* in weight space, and cannot find its way to better local minima, or better still, to the optimal weights that are the global minimum for error. Attempting to mitigate the problem of local minima is a major day-to-day concern of connectionist researchers, as well as being a focus of theoretical research. The problem of local minima can be reduced 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). But the adjustment of these parameters is often more a matter of judgment, experience, and guesswork than it is a product of theoretical analysis. 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.

A popular variation of the feed-forward network is the simple recurrent network (SRN; Elman, 1988, 1990) (see Figure 2.2). This network is essentially a standard feed-forward network equipped with an extra layer of so-called context units. At a particular time step an input pattern is propagated through the hidden-unit layer to the output layer (solid arrows). At the next time step the activation of the hidden-unit layer at the previous time step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows).[2] This means that the current state of the hidden units can influence the processing of subsequent inputs, providing a limited ability to deal with integrated sequences of input presented successively.

Whereas simple recurrent networks can be trained using the standard back-propagation learning algorithm, fully recurrent networks are trained using more complex learning algorithms, such as discrete back-propagation

Figure 2.2
Simple Recurrent Network



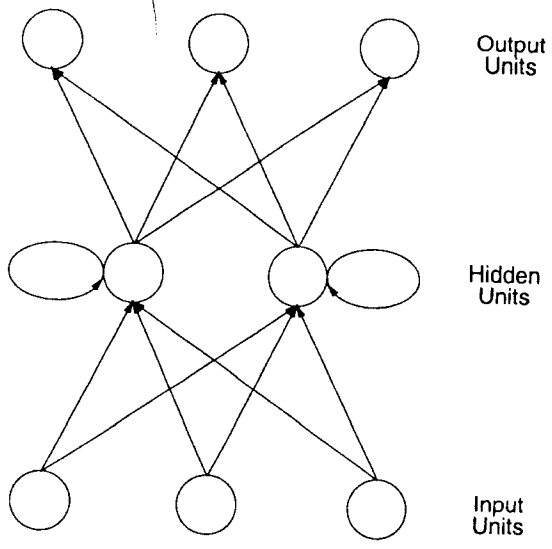
At a particular time step an input pattern is propagated through the hidden-unit layer to the output layer (solid arrows). At the next time step the activation of the hidden-unit layer at the previous time step is copied back to the context layer (dashed arrows) and paired with the current input (solid arrows).

through time (Williams & Peng, 1990) and continuous back-propagation (Pearlmutter, 1989). This type of network architecture is shown in Figure 2.3. Through the recurrent links (circular arrows), current activation can affect future activations similarly to the simple recurrent network, but in a more fine-grained manner and potentially reaching further back in time.

Another popular network architecture is the interactive activation network, shown in Figure 2.4. This type of network is completely prespecified (i.e., it does not learn). It consists of a sequence of unit layers. Units in the first layer typically encode fine-grained features of the input (e.g., visual or phonetic features). Units in the subsequent layers encode elements of increasingly higher levels of analyses (e.g., letters → words or phonemes → words). Units are connected using bidirectional links that can be either excitatory (arrows) or inhibitory (filled circles). This style of connectivity allows for activation to flow both bottom-up and top-down, reinforcing mutually consistent states of affairs and inhibiting mutually inconsistent states of affair.

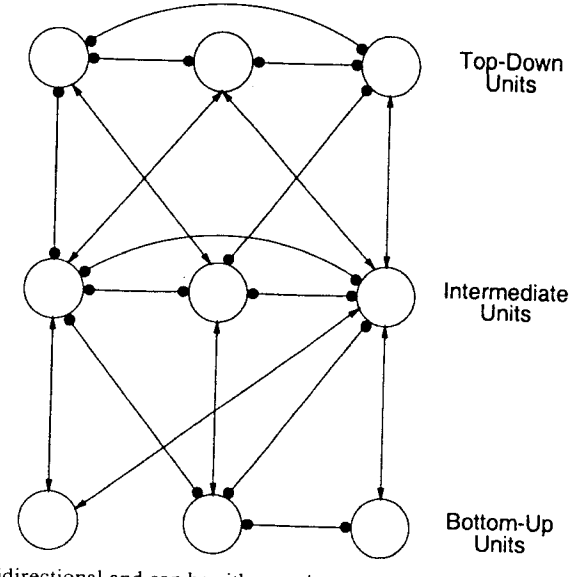
The behavior of individual units in interactive activation networks is somewhat more complex than in the network architectures we have described so far, because it depends not only on the current input but also on the previous

Figure 2.3
Fully Recurrent Network



Recurrent links (circular arrows) allow activation at the current time step to affect activations for many future time steps.

Figure 2.4
Interactive Activation Network



The links are bidirectional and can be either excitatory (arrows) or inhibitory (filled circles). Activation in this network flows both bottom-up and top-down.

level of activity of the unit. 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 outputs of the units that feed into 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).

While this behavior 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; small inputs lead to activations levels close to 0, while large inputs lead to activation values which tend to be 1 or −1. If we think of a unit as a feature detector, then an activation level near 1 corresponds to a high level of confidence that the feature is present; an activation level near −1 corresponds to a high level of confidence that it is not.

With respect to the relationship between connectionist models and the brain, it is important to note that none of the connectionist architectures that we have described amount to realistic models of brain function (see, e.g., Sejnowski, 1986). They are unrealistic at the level of individual processing units, where the models not only drastically oversimplify, but knowingly falsify, many aspects of the function of real neurons, and in terms of the structure of the connectionist 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 the areas of cognition in which few biological constraints are available, most notably language, researchers have concentrated on developing connectionist models with the goal of accurately modeling human behavior. They therefore take their data from cognitive psychology, linguistics, and cognitive neuropsychology, rather than from neuroscience. Thus, connectionist research on language appears to stand in direct competition with symbolic models of language processing.

As noted earlier, the relative merits of connectionist and symbolic models of language are hotly debated. But should they be viewed as in opposition at all? After 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. But symbolic theorists 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 is a crude approximation). If this is right, then connectionist modeling 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; Marcus, 1998; Pinker & Prince, 1988) typically assume that it implies that symbolic modeling 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 influences 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. Part II of this book will illustrate a variety of contemporary viewpoints on the relationship between connectionist and symbolic theories of language.

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

SPEECH PROCESSING

Speech processing in its broadest sense encompasses a broad range of cognitive processes, from those involved in low-level acoustical analysis to those involved in semantic and pragmatic interpretation of utterances. Here we shall focus much more narrowly, on the processes involved in segmenting and recognizing spoken words from input that is represented in a linguistic form (e.g., as sequences of phonetic features or phonemes). Thus, we will not be concerned with connectionist research on the enormously complex issues involved in dealing with the complexity, variability, and noisiness of acoustic representations of speech (see, e.g., Salmela, Lehtokangas, & Saarinen, 1999, for a typical application of connectionist methods to speech technology). We also shall not deal with higher-level aspects of linguistic processing. Nonetheless, as we shall see, even given these restrictions, the problem of understanding human speech processing is still formidable.

Naïvely, we might imagine that the speech processor has to do two jobs, one after the other. First, it has to segment speech input into units corresponding to words (i.e., it has to find word boundaries); second, it has to recognize each word. But on reflection, this viewpoint seems potentially problematic, because it is not clear how the speech processor can determine where the word boundaries are until the words are recognized. And conversely, word recognition itself seems to presuppose knowing which chunk of speech material corresponds to a potential word. Thus, segmentation and recognition appear to stand in a chicken-and-egg relationship—each process seems to depend on the other.

One approach to resolving the paradox is to assume that segmentation and recognition are two aspects of a single process, that tentative hypoth-

eses about each issue are developed and tested simultaneously, and mutually consistent hypotheses are reinforced. A second approach is to suppose that there are segmentation cues in the input that are used to give at least better-than-chance indications of what segments may correspond to identifiable words. So the question is this: Does speech processing involve dedicated segmentation strategies prior to word recognition?

Developmental considerations suggest that there may be specialized segmentation methods. The infant, initially knowing no words, seems constrained to segment speech input using some method not requiring word recognition. Moreover, infant studies have shown that prelinguistic infants may use such methods, and are sensitive to a variety of information that is available in the speech stream and potentially useful for segmentation, such as phonotactics and lexical stress (Jusczyk, 1997).

Connectionist models have begun to address questions of how effective different kinds of segmentation cues might be. For example, Cairns, Shillcock, Chater, and Levy (1997) explore a model of segmentation based on predictability. They note that language is less predictable across, rather than between, words. They trained a recurrent network on a large corpus of phonologically transcribed conversational speech, represented as a sequence of bundles of binary phonetic features. The network was trained to predict the next bundle of features along with the previous and current feature bundles, based on the current input material. Where prediction error was large, it was assumed that a word boundary had been encountered. This model captured some aspects of human segmentation performance. For example, it spontaneously learned to pay attention to patterns of strong and weak syllables as a segmentation cue. However it was able to reliably predict only a relatively small proportion of word boundaries, indicating that other cues also need to be exploited. While the Cairns et al. model uses just a single cue to segmentation, Christiansen, Allen, and Seidenberg (1998) showed how multiple, partial constraints on segmentation could yield much better segmentation performance. They trained an SRN to integrate sets of phonetic features with information about lexical stress (strong or weak) and utterance boundary information (encoded as a binary unit) derived from a corpus of child-directed speech. The network was trained to predict the appropriate values of these three cues for the next segment. After training, the network was able to integrate the input such that it would activate the boundary unit not only at utterance boundaries, but also at word boundaries inside utterances. The network was thus able to generalize patterns of cue information that occurred at the end of utterances to cases where the same patterns occurred within an utterance. This model performed well on the word-segmentation task while capturing additional aspects of infant segmentation, such as the bias toward the dominant trochaic (strong-weak) stress pattern in English, the ability to distinguish between phonotactically

legal and illegal novel words, and having segmentation errors being constrained by English phonotactics.

This model shows how integrating multiple segmentation cues can lead to good segmentation performance. To what extent does it provide a model of how infants process speech? Christiansen, Conway, and Curtin (2000) used the trained model, without any additional modifications, to fit recent infant data. These data are of particular interest, because they have been claimed to be incompatible with a purely connectionist approach to language processing, and to require the language processor to use "algebraic" or symbolic rules (Marcus, Vijayan, Rao, & Vishton, 1999). Marcus et al. habituated infants on syllable sequences that followed either an AAB or ABB pattern (e.g., *le-le-je* versus *le-je-je*). The infants were then presented with sequences of novel syllables, either consistent or inconsistent with the habituation pattern, and showed a preference for the inconsistent items. Christiansen et al. suggested that statistical knowledge acquired in the context of learning to segment fluent speech provided the basis for these results, in much the same way as knowledge acquired in the process of learning to read can be used to perform experimental tasks such as lexical decision. Their simulation closely replicated the experimental conditions, using the same number of habituation and test trials as in the original experiment (no repeated training epochs) and one network for each infant. Analyses of the model's segmentation performance revealed that the model was significantly better at segmenting out the syllables in the inconsistent items. This makes the inconsistent items more salient and therefore explains why the infants preferred these to the consistent items. Thus, Christiansen et al.'s results challenge the claim that the Marcus et al. infant data necessarily require that the infant's language-processing system is using algebraic rules. Moreover, these infant data provide an unexpected source of evidence for the Christiansen et al. model, viewed as a model of infant segmentation.

Segmentation cues are potentially important in guiding the process of word recognition. But even if such cues are exploited very effectively, segmentation cues alone can achieve only limited results. A definitive segmentation of speech can only occur after word recognition has occurred. Speech is frequently locally ambiguous: To use an oft quoted example, it is difficult to distinguish "recognize speech" from "wreck a nice beach" when these phrases are spoken fluently. These interpretations correspond to very different segmentations of the input. It is therefore clear that bottom-up segmentation cues alone will not always segment the speech stream into words reliably. In such cases of local ambiguity, a decisive segmentation of the input can only be achieved when the speaker has recognized which words have been said. This theoretical observation ties in with empirical evidence that strongly indicates that during word recognition in adulthood multiple candidate words are activated, even if these correspond to different segmen-

tation of the input. For example, Gow and Gordon (1995) found that adult listeners hearing sentences involving a sequence (e.g., *two lips*) that could also be a single word (*tulips*) showed speeded processing of an associate of the second word (*kiss*) and to an associate of the longer word (*flower*), indicating that the two conflicting segmentations were simultaneously entertained. This would not occur if a complete segmentation of the input occurred before word recognition was attempted. On the other hand, it is not clear how these data generalize to word segmentation and recognition in infancy before any comprehensive vocabulary has been established. How segmentation and recognition develop into the kind of integrated system evidenced by the Gow and Gordon data remains a matter for future research.

Gow and Gordon's (1995) result also suggests that word recognition itself may be a matter of competition between multiple activated word representations, where the activation of the word depends on the degree of match between the word and the speech input. Indeed, many studies point toward this conclusion, from a range of experimental paradigms. Such competition is typically implemented in connectionist networks by a localist code for words (the activation of a single unit represents the strength of evidence for that word, with inhibitory connections between word units). Thus, when an isolated word is identified, a "cohort" of words consistent with that input is activated; as more of the word is heard, this cohort is rapidly reduced, perhaps to a single item.

While competition at the word level has been widely assumed, considerable theoretical dispute has occurred over the nature of the interaction between different levels of mental representation. Bottom-up (or "data-driven") models are those in which less abstract levels of linguistic representation feed into, but are not modified by, more abstract levels (e.g., the phoneme level feeds to the word level, but not the reverse). We note, however, that this does not prevent these models from taking advantage of suprasegmental information, such as in the inclusion of lexical stress in the Christiansen et al. (2000) segmentation model, provided that this information is available in a purely bottom-up fashion (i.e., no lexical-level feedback). Interactive (also "conceptually-driven" or top-down) models allow a two-way flow of information between levels of representation. Figures 2.1 and 2.4 provide abstract illustrations of the differences in information flow between the two types of models. Note that bottom-up models allow information to flow through the network in one direction only, whereas interactive models allow information to flow in both directions.

The bottom-up versus interactive debate rages in all areas of language processing, and also in perception and motor control (e.g., Bruner, 1957; Fodor, 1983; Marr, 1982; Neisser, 1967). Here we focus on putative interactions between information at the phonemic and the lexical levels in word recognition (i.e., between phonemes and words), where experimental work and connectionist modeling has been intense.

The most obvious rationale for presuming that there is top-down information flow from the lexical to the phoneme level stems from the effects of lexical context on phoneme identification. For example, Ganong (1980) showed that the identification of a syllable-initial speech sound, constructed to be between a /g/ and a /k/, was influenced by lexical knowledge. This intermediate sound was predominantly heard as a /k/ if the rest of the word was -iss (*kiss* was favored over *giss*), but heard as /g/ if the rest of the word was -ift (*gift* was favored over *kift*).

The early and very influential TRACE model of speech perception (McClelland & Elman, 1986) attempts to explain data of this kind from an interactive viewpoint. The model employs the standard interactive activation network architecture already described, 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 centered over every three time slices. The phoneme detector centered 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. TRACE captures the Ganong 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 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).

TRACE provides an impressive demonstration that context effects can indeed be modeled from an interactive viewpoint. But context effects on phoneme recognition can also be explained in purely bottom-up terms. If a person's decisions about phoneme identity depend on both the phonemic and lexical levels, then phoneme identification may be lexically influenced, even though there need be no feedback from the lexical to the phoneme level. For example, the Ganong effect might be explained by assuming that

the phoneme identification of an initial consonant that is ambiguous between /g/ and /k/ is directly influenced by the lexical level. Thus, if *gift* is recognized at the lexical level, this will influence the participant to respond that the initial phoneme was a /g/, but if *kiss* is recognized, this will influence the participant to respond that the initial phoneme was a /k/.

A substantial experimental literature has attempted to distinguish TRACE from bottom-up models, indicating the importance of connectionist modeling in inspiring experimental research. One line of attack was that the interactive character of TRACE causes word-level context to have too abrupt an effect in modulating phoneme perception. Massaro (1989) had people listen to phonemes on the continuum between /r/ and /l/ in syllables in which they were immediately preceded by /t/ or /s/. In normal English these preceding phonemes are highly informative about the ambiguous /r/–/l/. With an initial /t/, the next item must be an /r/, because /t/ followed by /l/ is not permissible according to the phonotactics of English (i.e., the constraints on legal phoneme sequences). Conversely, with an initial /s/, the next item must be an /l/, because /s/ followed by /r/ is not phonotactically permissible.

When TRACE is applied to these stimuli, there is a relatively abrupt switch from the one interpretation of the ambiguous phoneme to the other. For example, in the context of an /s/, TRACE adopts the context-appropriate interpretation that the ambiguous phoneme is an /l/ until the ambiguous phoneme is perceptually very strongly biased toward /r/, at which point the model switches sharply to the opposite interpretation. The opposite pattern was observed in the context of a /t/. But this sharp crossover was not consistent with the human data that Massaro (1989) collected. Instead, people required less perceptual bias to override word-level context, but the transition between judging the ambiguous phoneme to be an /r/ or an /l/ was also much more gradual over the /r/–/l/ continuum. Massaro concluded that the interactive activation approach was flawed, because it allowed more distortion of the perceptual stimulus by contextual factors than is actually observed. Massaro further showed that a bottom-up model, based on his Fuzzy Logic Model of Perception (FLMP), could account for the empirical data more accurately. FLMP involves a simple linear combination of cues from the perceptual stimulus and surrounding context: It can be viewed as equivalent to a single-layer connectionist network (with noise added, so that the network is merely biased toward producing one response more than the other, rather than producing the same output deterministically).

McClelland (1991) showed, however, that adding noise to the units in TRACE could also produce a more graded pattern of responding. Intuitively, adding noise to any decision-making system will inevitable blur the boundary between the inputs that typically give rise to one decision and the inputs that typically give rise to another. Massaro and Cohen (1991) responded that there are other aspects of the data that McClelland's revised interactive model does not capture. On the other hand, McClelland's model

provides a much more detailed computational mechanism than is provided by FLMP, so a fair comparison between the two is not straightforward.

Another potential difficulty for TRACE in relation to data on phoneme perception is that the influence of lexical factors on phoneme perception appears to be quite variable. In some experiments substantial lexical influences on phoneme judgments are observed; but in others, often differing only slightly, the effects disappear. For example, Cutler, Mehler, Norris, and Segui (1987) found lexical effects when participants were asked to monitor for initial phonemes of monosyllabic targets only when the filler items varied in syllabic length, and McQueen (1991) found lexical influences on the categorization of ambiguous word-final fricatives (on the continuum between /s/ and /ʃ/) only when the stimuli were perceptually degraded by low-pass filtering them at 3kHz. These and other studies (see Pitt & Samuel, 1993, for a review) present a confusing picture for any theoretical account.

Opponents of the interactive character of TRACE argue that some of these data can be understood by assuming a bottom-up model in which phoneme judgments are jointly influenced by a phonemic level of representation and a lexical level of representation. According to this view, the direct influence of the lexical level of representation on phoneme judgments (although not on the level of phoneme representation) is the source of context effects, and these effects can be turned on and off to the degree that the task demands encourage participants to attend to the lexical level. Thus, if the filler and target items are monosyllabic, they may become monotonous and discourage attention at the lexical level; if targets are perceptually degraded, this may encourage attention to the lexical level, because perceptual representations are not sufficiently reliable to be used alone. Thus, critics of TRACE have argued that it is limited by having only one "route" by which phonetic judgments can be made—this route depending directly on the representations at the phoneme level. By contrast, in order to capture context effects, bottom-up models necessarily allow two routes that can influence phonemic judgments, via phonemic and lexical representations. Various models, both nonconnectionist (Cutler & Norris, 1979) and connectionist (Norris, McQueen, & Cutler, in press), have been proposed in opposition to TRACE. These models exploit two routes, and hence allow for the possibility of "attentional" switching between them.

Despite the considerable empirical and theoretical interest that these issues have attracted, it seems unlikely that the instability of lexical influences will be decisive in determining whether speech perception is viewed as bottom-up or interactive. This is because the simple expedient of allowing that phonemic judgments can depend on the activations of both the phonemic and lexical level in TRACE immediately give the interactive account precisely the same explanatory latitude as bottom-up models. There is no reason why interactive models cannot also assume that a variety of levels of representations may be drawn upon in performing a specific task. Nonethe-

less, although this theoretical move is entirely viable for advocates of the interactive position, it is unattractive on the grounds of parsimony. This is because the resulting model would have two routes by which contexts effects could arise, one due to the direct influence of high-level representations on task performance (specifically, the influence of word-level representations on phoneme judgment tasks), and the other due to the indirect, top-down impact of higher levels on lower levels (specifically, the top-down links from the word to the phoneme level). If a bottom-up account can explain the same data using just one mechanism—the direct influence of higher-level representations on phoneme judgments—then this viewpoint has a considerable advantage in terms of parsimony.

One key experimental result (Elman & McClelland, 1988), derived as a novel prediction from TRACE, appeared to be particularly persuasive evidence against bottom-up connectionist models. 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." Listeners should therefore adjust their category boundaries depending on the phonemic context. Experiments confirm that people do indeed exhibit this "compensation for coarticulation" (CFC; Mann and Repp, 1980). For example, given a series of synthetically produced tokens between /t/ and /k/, listeners move the category boundary toward the /t/ following a /s/ and toward the /k/ following a /ʃ/. 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/ influences a word-initial phoneme ambiguous between /t/ and /k/ to be heard as a /k/ (as in *Christmas capes*). If lexical-level representations feed back onto phoneme-level representations, the compensation of the /k/ 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 should proceed as normal). Elman and McClelland confirmed TRACE's prediction experimentally. Recognition of the phoneme at the start of the second word was apparently influenced by CFC, as if the word-final phoneme in the first word had been "restored" by lexical influence.

Surprisingly, bottom-up connectionist models can also capture these results. Norris (1993) provided a small-scale demonstration, training an SRN to map phonetic input onto phoneme output, for a small (twelve-word vocabulary) artificial language. When the network received phonetic input with an ambiguous first word-final phoneme and ambiguous initial segments of the second word, an analog of CFC was observed. 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 (1988). But the explanation for this pattern of results cannot be top-down influence from word units, because there are no word units. Moreover, Cairns, Shillcock, Chater, and Levy (1995) scaled up these results using a similar network trained on phonologically transcribed conversational English. How can an autonomous computational model, where there is no lexical influence on phoneme processing, mimic the apparent influence of word recognition on coarticulation? Cairns et al. argued that sequential dependencies between the phoneme sequences in spoken English can often "mimic" lexical influence. The idea is that the identification of the word-final ambiguous phoneme favored by the word level is also typically favored by transitional probability statistics across phonemes. Analyzing statistical regularities in the phoneme sequences in a large corpus of conversational English, Cairns et al. showed that this explanation applies to Elman and McClelland's experimental stimuli. If these transitional probabilities have been learned by the speech processor, then previous phonemic context might support the "restoration" of the ambiguous word-final phoneme, with no reference to the word in which it is contained. Pitt and McQueen (1998) tested between these two explanations experimentally. They carefully controlled for transitional probabilities across phonemes, and reran a version of Elman and McClelland's experiment: Compensation for coarticulation was eliminated. Moreover, when transitional probabilities are manipulated in nonword contexts, compensation for coarticulation effects were observed. This pattern of results suggests that compensation for coarticulation is not driven by top-down lexical influence, but by phoneme-level statistical regularities.

Against this, Samuel (1996) argues that the precise pattern of phoneme restoration does indicate the existence of small but discernible top-down effects. He conducted a statistical analysis of people's ability to discriminate whether a phoneme has been replaced by a noise in a word or nonword context from the case where the phoneme and noise are both present. The logic is that to the extent that top-down factors "restore" the missing phoneme, it should be difficult to tell whether or not the phoneme is actually present, and hence people's discrimination between the two cases should be poorer. Hence, phoneme present–absent discrimination should be poorer in word contexts than for nonword contexts, because top-down factors should be stronger. This prediction was confirmed experimentally (Samuel, 1996). However, this pattern of data also follows from bottom-up models, to the extent that the judgment concerning whether the phoneme is present is determined not only by phonological but also lexical representations. Accordingly, in a word context, judgments will be potentially biased by the word-level representation signaling that the missing phoneme is present (because a word in which that phoneme normally occurs has been recognized). This line of evidence, therefore, does not seem to provide a strong way of

distinguishing between bottom-up and top-down accounts, and there are connectionist models compatible with Samuel's data that operate in each way (McClelland & Elman, 1986; Norris et al., in press).

Another recent study by Samuel may pose a more difficult challenge to bottom-up accounts. Samuel (1997) uses the fact that hearers adapt their classification of speech continua, such as the continuum between /b/ and /d/, after hearing a word beginning with a speech sound at one end of the continuum. For example, after hearing *bird*, a hearer's category boundary shifts toward /b/ on the /b/–/d/ continuum. The logic of Samuel's study was to investigate whether adaptation can occur to a word in which the key initial phoneme is perceptually restored, rather than actually presented. Thus, Samuel presents words in which the initial phoneme (/b/ or /d/) is replaced by noise, as in a typical phoneme-restoration study. As expected, he found that participants restored the "missing" phoneme, even though it was not present. But crucially, he also found that these words did indeed produce an adaptation effect with respect to the categorization of ambiguous phonemes on the continuum from /b/ to /d/. The effect did not occur where the deleted phonemes were replaced by silence rather than a burst of noise, and hence there was no perceived phoneme restoration. Samuel argues that this pattern of results indicates that the phoneme representations are being affected by the lexical level, and that this leads to the adaptation. Norris et al. (in press) note that it remains to be shown that the adaptation effect is itself mediated by the phoneme level: If, for example, adaptation effects for phonemes could be caused directly by the lexical level, then this would provide a possible account of Samuel's data. Nonetheless, if Samuel's result does prove to be robust, it could be extremely difficult for bottom-up accounts to deal with, except by using rather ad hoc explanations.

Finally, additional evidence for the ability of bottom-up models to accommodate apparently lexical effects on speech processing was provided by Gaskell, Hare, and Marslen-Wilson (1995). They trained an SRN model to map a systematically altered featural representation of speech onto a canonical representation of the same speech, and found that the network showed evidence of lexical abstraction (i.e., tolerating systematic phonetic variation, but not random change). More recently, Gaskell and Marslen-Wilson (1997) have added a new dimension to the debate, presenting an SRN network in which sequentially presented phonetic inputs for each word were mapped onto corresponding distributed representations of phonological surface form and semantics. Based on the ability of the network to model the integration of partial cues to phonetic identity and the time course of lexical access, they suggested that distributed models may provide a better explanation of speech perception than their localist counterparts (e.g., TRACE). An important challenge for such distributed models is to accommodate the simultaneous activation of multiple lexical candidates necessitated by the temporal ambiguity of the speech input (e.g., /kæp/ could be the beginning

of both *captain* and *captive*). The coactivation of several lexical candidates in a distributed model results in a semantic "blend" vector. Through statistical analyses of these vectors, Gaskell and Marslen-Wilson (Chapter 3, this volume) investigate the properties of such semantic blends, and apply the results to explain some recent empirical speech-perception data.

The theoretical debate concerning segmentation and word recognition has been profoundly influenced by connectionist psycholinguistics. We have considered various streams of research arising out of the TRACE model of speech perception to illustrate the interplay between connectionist modeling and experimental studies, and there are many other important areas of research we have not considered for lack of space (e.g., recent work on an apparent interaction between phonetic mismatch and lexical status, which has triggered a subtle and important strand of research; Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, in press). Connectionist models are now the dominant style of computational account, even for advocates of very different positions (as we have seen in relation to the bottom-up–interactive debate). Attempts to test between the predictions of competing models have generated experimental advances that have in turn informed how models develop. However, this progress has yet not resulted in a resolution of the fundamental debate between proponents of bottom-up and interactive approaches to speech processing, though Norris et al. (in press) may provide some advantage for the opponents of top-down lexical effects.

Overall, these studies indicate how connectionist models of speech processing have been able to make good contact with detailed psycholinguistic data and been important in motivating experimental work. Input representativeness is also generally good, with models being trained on large lexicons and sometimes corpora of natural speech. Task veridicality may perhaps be questioned, however, by the use of abstract representations of the input (e.g., phonetic or phonological representations) that may not be computed by the listener (Marslen-Wilson & Warren, 1994), and that also bypass the deep problems involved in handling the physical variability of natural speech.

MORPHOLOGY

One of the connectionist models that has created the most controversy is Rumelhart and McClelland's (1986a) model of the learning of the English past tense. 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. 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 rulelike behavior? Rumelhart and McClelland's (1986a) model of the acquisition of the past tense in English was presented as an affirmative answer to

this question. The English past tense is an interesting test case because children very roughly appear to exhibit U-shaped learning, traditionally characterized as having three stages. 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*. It is important that children now show evidence of rulelike behavior. They are able to conjugate nonwords, generating *jicked* as the past tense of *jick*, and they start to overgeneralize irregular verbs, even the ones they got right in stage one; 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 that children learn to use a rule-based route for dealing with regulars as well as nonwords and a memorization route for handling irregulars. But how can such seemingly dual-route behavior be accommodated by a single mechanism employing just a single route?

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. 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 that takes phonological representations of root forms and transforms them into sets of phonetic feature triples, termed wickelfeatures (after Wickelgren, 1969, who employed triples in modeling memory for sequential material).[3] In order to promote generalization, 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 pattern-associator network, which is the core of the model, 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 "the rich get richer" competitive approach). The end result of this competition is a set of more or less nonoverlapping wickelphones that correspond to as many as possible of the wickelfeatures in the input to the decoder network.

By employing a particular training regime, Rumelhart and McClelland (1986a) were able to obtain the U-shaped learning profile characteristic of children's acquisition of the English past tense. First, the network was trained on a set of 10 high-frequency verbs (8 irregular and 2 regular) for 10 epochs. At this point the network 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 network was trained for an additional 190 epochs. Early on during this period of training the net behaved as children at acquisition stage 2: The network tended to regularize irregulars while getting regulars correct. At the end of the 190 epochs, network behavior resembled that of children in stage 3 of the past-tense acquisition process, exhibiting an almost perfect performance on the 420 verbs. The network 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 generalization from the 420 verbs in the training set to a separate test set consisting of 86 low-frequency verbs (of which just over 80% were regular); 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 /d/ preceded by an unstressed vowel (schwa) with verb stems ending with a *t* or a *d*).

The merits and inadequacies of the Rumelhart and McClelland (1986a) past-tense model has been the focus of much debate, originating with Pinker and Prince's (1988) detailed criticism (and to a lesser extent 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 & Seidenberg, 1992; Daugherty, MacDonald, Petersen, & Seidenberg, 1993; MacWhinney & Leinbach, 1991; Seidenberg, 1992). Here we focus on the most influential aspects of the debate.

The use of wickelphones–wickelfeature representations has been the subject of much criticism (e.g., 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 diachronic change, Hare & Elman, 1995) therefore use a position-specific phonological representation in which vowels and consonants are defined in terms of sets of phonetic features. Another, more damaging criticism of the single-route approach is that the U-shaped pattern of behavior observed in the model during learning essentially appears 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 explanation of U-shaped learning suggested by 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 feed-forward network with a hidden-unit layer on a vocabulary of artificial verb stems and past-tense forms, patterned by regularities of the English past tense. They held the size of the vocabulary used in training constant at 500 verbs. They found that the network 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 behavior. For example, given a training set with a type and token frequency reflecting that of English verbs, the network was able to simulate a number of subregularities between the phonological form of a verb stem and its past tense form (e.g., *sleep* → *slept*, *keep* → *kept*).[4] 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 network (compared with a similarly configured network trained on the same vocabulary in Plunkett & Marchman, 1991). This approach also suggested that a critical mass of verbs is needed before a change from rote learning (memorization) to system building (rulelike generalization behavior) 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 fifty words (e.g., Bates, Bretherton, & Snyder, 1988). Plunkett and Juola (Chapter 4, this volume) find a similar critical-mass effect in their model of English noun and verb morphology. They analyzed the developmental trajectory of a feed-forward network trained to produce the plural form for 2,280 nouns and the past tense form for 946 verbs. The model exhibited patterns of U-shaped development for both nouns and verbs (with noun inflections acquired earlier than verb inflections), and also demonstrated a strong tendency to regularize deverbal nouns and denominal verbs.

Another criticism of the connectionist models of past-tense acquisition is that they may be 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 generalization ability. In English, most verbs are regular—that is, 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 memorization of the irregular past tenses in terms of a number of narrow phonological subcategories (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 showed that the default generalization in Rumelhart and McClelland's (1986a) model was dependent on a similar frequency distribution in the training set. They furthermore contended that no connectionist model can accommodate default generalization for a class of words that have both low type frequency and low token frequency. The default inflection of plural nouns in German appears 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, & Pinker, 1995). If true, such lack of cross-linguistic validity would render connectionist models of past-tense acquisition obsolete.

However, recent connectionist work has addressed the issue of minority default mappings with some success. Daugherty and Hare (1993) trained a feed-forward 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 A.D. 870). The training set consisted of five classes of irregular verbs plus one class of regular verbs, each class containing twenty-five words (each represented once in the training set). Thus, words taking the default generalization *-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 behavior even when faced with a low-frequency default class. Indeed, it appears that generalization in connectionist networks may not be strictly dependent on similarity to known items. Daugherty and Hare's results show that if the nondefault (irregular) classes have a sufficient degree of internal structure, default generalization 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 feed-forward 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 characterized by a system of subregularities dependent on the phonological shape of the noun stem. In contrast, a minority of nouns take the Sound Plural inflection that 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 generalization without relying on large word classes or direct similarity.

Finally, rulelike and frequency-independent default generalization may not be as pressing a problem for connectionist models as Clahsen et al. (1993) and Marcus et al. (1995) claim. Reanalyzing data concerning German noun inflection (in combination with additional data from Arabic and Hausa), Bybee (1995) showed that default generalization is sensitive to type frequency and does not seem to be entirely rulelike. This pattern may fit better with the kind of default generalization in connectionist nets rather than the rigid defaults of symbolic models. Moreover, Hahn and Nakisa (in press) outline problems for the dual-route approach. They compared connectionist and other implementations of rule and memorization routes against a single memorization route on a comprehensive sample of German nouns and found that performance was consistently superior when the rule route was not used.

The issue of whether humans employ a single, connectionist-style mechanism for rulelike 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 concede that some kind of associative connectionist mechanism may explain the complex patterns found in the irregular cases (e.g., Pinker, 1991). The controversial question is whether a single connectionist mechanism can simultaneously account for both regular and irregular cases, or whether regular cases can only be generated by a distinct route involving (perhaps necessarily symbolic) rules.

Most of the connectionist models of morphology only make contact with secondary empirical data. Many of the models suffer from low task veridicality because they are trained to map verb stems to past-tense forms (e.g., Plunkett & Marchman, 1991, 1993; Rumelhart & McClelland, 1986a; but see, e.g., Hoeffner, 1997, for an exception), a task unlikely to be relevant to children's language acquisition. However, rule-based morphology models (e.g., Pinker, 1991) also involve stem to past-tense mappings as the connectionist models, and thus suffer from the same low task veridicality. Input representativeness, on the other hand, is reasonable; Plunkett and Joula (Chapter 4, this volume) provide a good example in this respect. The future is likely to bring further connectionist modeling 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 processing can be accounted for without the explicit representation of rules also plays an important part in connectionist modeling of sentence processing.

SENTENCE PROCESSING

Syntactic processing is arguably the area of natural language that 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? Much of the recent connectionist research on sentence processing aims to show that it can.

Sentence processing provides a considerable challenge for connectionist research. In view of the difficulty of the problem, much early work "hand-coded" symbolic structures directly into the network architecture; starting with Small, Cottrell, and Shastri's (1982) first attempt at connectionist parsing, followed by Reilly's (1984) connectionist account of anaphor resolution, and later, for example, by Fanty's (1985) connectionist context-free parser, Selman and Hirst's (1985) modeling of context-free parsing using simulated annealing, Waltz and Pollack's (1985) interactive model of parsing (and interpretation), McClelland and Kawamoto's (1986) connectionist model of case-role assignment, and, more recently, Miyata, Smolensky, and Legendre's (1993) structure-sensitive processing of syntactic structure using tensor representations (Smolensky, 1990) as well as Kwasny and Faisal's (1990) deterministic connectionist parser. Such connectionist reimplementations 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). But most connectionist research has a larger goal: to provide alternative accounts of sentence processing in which networks learn to form and use structured representations rather than simply implement symbolic representations and processes.

Two classes of models potentially provide such alternatives. Both classes of model learn to process language from experience, rather than implementing a prespecified set of symbolic rules. The less ambitious class presup-

poses 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 to 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's (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 thereof) that 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 (e.g., Brill, Magerman, Marcus, & Santorini, 1990; Charniak, 1993; Jelinek, Lafferty, & Mercer, 1990), in which probabilities of grammar rules in a prespecified context-free grammar are learned from a corpus of parsed sentences. Another approach within this class of connectionist models—sometimes referred to as "structured connectionism"—involves the construction of a modular system of networks, each of which is trained to acquire different aspects of syntactic processing. For example, Miikkulainen's (1996) system consists of three different networks: one trained to map words onto case-role assignments, another trained to function as a stack, and a third trained to segment the input into constituent-like units. Although the model displays complex syntactic abilities, the basis for these abilities and their generalization to novel sentence structures derive from the configuration of the stack network combined with the modular architecture of the system, rather than being discovered by the model.

The second, more ambitious class of models, which includes Christiansen and Chater (Chapter 5, this volume) as well as Tabor and Tanenhaus (Chapter 6, this volume), attempts the much harder task of learning syntactic structure from sequences of words, with no explicit prior assumptions about the particular form of the grammar. These models have only recently begun to provide accounts for empirical sentence-processing phenomena. This may explain why the more ambitious 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 discussed later (Seidenberg & McClelland, 1989). Nevertheless, these models may potentially have a great impact on the psychology of language because they 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).

The most influential approach of this kind is due to Elman (1991, 1993), who trained an SRN to predict the next input word for sentences generated by a small context-free grammar. This grammar involved subject noun–verb agreement, variations in verb argument structure (i.e., intransitive, transitive, optionally transitive), and subject and object relative clauses (allowing multiple embeddings with complex long-distance dependencies). Elman's simulations suggested that an SRN can acquire some of the grammatical regularities underlying a grammar. In addition, the SRN showed some similarities with human behavior on center-embedded structures (Weckerly & Elman, 1992). Christiansen (1994, 2000) extended this work, using more complex grammars involving prenominal genitives, prepositional modifications of noun phrases, noun-phrase conjunctions, and sentential complements, in addition to the grammatical features used by Elman. One of the grammars, moreover, incorporated cross-dependencies, a weakly context-sensitive structure found in Dutch and Swiss-German. Christiansen found that SRNs could learn these more complex grammars, and, moreover, that they exhibit the same qualitative processing difficulties as humans do on similar constructions (see also Christiansen & Chater, Chapter 5, this volume). The nets moreover showed sophisticated generalization abilities, overriding local word cooccurrence statistics while complying with structural constraints at the constituent level (Christiansen & Chater, 1994).

Current models of syntax typically use "toy" fragments of grammar and small vocabularies. Aside from raising questions about how well the results will scale up, this makes it difficult to provide detailed fits with empirical data. Nonetheless, some attempts have recently been made toward fitting existing data and deriving new empirical predictions from the models. For example, Tabor, Juliano, and Tanenhaus (1997) provide a two-component model of ambiguity resolution, combining an SRN with a "gravitational" mechanism. The SRN was trained in the usual way on sentences derived from a grammar. After training, SRN hidden-unit representations for individual words were placed in the gravitational mechanism, which was then allowed to settle into a stable state. Settling times were then mapped onto word-reading times. Using their two-component model, Tabor et al. were able to fit data from several experiments concerning the interaction of lexical and structural constraints on the resolution of temporary syntactic ambiguities (i.e., garden-path effects) in sentence comprehension. Tabor and Tanenhaus (Chapter 6, this volume) extend the two-component model to account for empirical findings reflecting the influence of semantic role expectations on syntactic-ambiguity resolution in sentence processing (McRae, Spivey-Knowlton, & Tanenhaus, 1998).

In a different strand of research concerned with relating connectionist networks to psycholinguistic results, Christiansen and Chater (1999) developed a measure of grammatical prediction error (GPE) that allowed network output to be mapped onto human performance data. GPE scores are

computed for each word in a sentence and reflect the processing difficulties that a network is experiencing at a given point in a sentence. Averaging GPE across a whole sentence, Christiansen (2000; Christiansen & Chater, Chapter 5, this volume) fitted human data concerning the greater perceived difficulty associated with center-embedding in German compared to cross-serial dependencies in Dutch (Bach, Brown, & Marslen-Wilson, 1986). Christiansen was able to derive novel predictions concerning other types of recursive constructions, and these predictions were later confirmed experimentally (Christiansen & MacDonald, 2000). MacDonald and Christiansen (in press) mapped single-word GPE scores directly onto reading times, providing an experience-based account for human data concerning the differential processing of singly center-embedded subject and object relative clauses in human participants with different levels of reading comprehension ability.

Some headway has also been made in accounting for data concerning the effects of aphasia on grammaticality judgments. Allen and Seidenberg (1999) trained a recurrent network to mutually associate two input sequences: a sequence of word forms and a corresponding sequence of word meanings. The network was able to learn a small artificial language successfully; it was able to regenerate the word forms from the meanings and vice versa. Allen and Seidenberg simulated grammaticality judgments by testing how well the network could recreate a given input sequence, allowing activation to flow from the provided input forms to meaning and then back again. Ungrammatical sentences were recreated less accurately than grammatical sentences, and the network was thus able to distinguish grammatical from ungrammatical sentences. They lesioned the network by removing 10 percent of the weights in the network. Grammaticality judgments were then elicited from the impaired network for ten different sentence types that Linebarger, Schwartz, and Saffran (1983) used in their study of aphasic grammaticality judgments. The network exhibited impaired performance on exactly the same three sentence types as the aphasic patients.

These simulation results suggest that recurrent networks may be viable models of sentence 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 generalization abilities that would be expected from models of language. Hadley (1994a) correctly pointed out that generalization in much connectionist research has not been viewed in a sophisticated fashion. The testing of generalization 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" generalization a network must be shown to learn a word in one syntactic position and then generalize 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 generalize successfully in learning syntactically structured material.

Christiansen and Chater (1994) addressed this challenge, providing a formalization of Hadley's original ideas as well as presenting evidence that connectionist models are able to attain strong generalization. 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 been presented with singular verbs following *boy*. Nonetheless, the network was able to correctly predict that a plural verb must follow *John and boy* as prescribed by the grammar. In addition, the network was still able to correctly predict a plural verb when a prepositional phrase was attached to *boy*, as in *John and boy from town*, providing even stronger evidence for strong generalization. This suggests that the SRN is able to make nonlocal generalizations 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. More recently, Christiansen (2000) demonstrated that the same SRN also was able to generalize appropriately when presented with completely novel words, such as *zorg*, in a noun-phrase conjunction by predominately activating the plural verbs.[5] 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 generalize 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 generalization abilities, ignoring local word cooccurrence constraints while appearing to comply with structural information at the constituent level. Additional evidence of strong generalization 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 generalization. When Redington, Chater, and Finch (1993) applied a method of distributional statistics (see also Finch & Chater, 1993; Redington, Chater, & Finch, 1998) to a corpus of child-directed speech (the CHILDES corpus collected by MacWhinney & Snow, 1985), they found that the syntactic

category of a nonsense word could be derived from a single occurrence of that word in the training corpus. This indicates that strong generalization may be learnable through the kind of bottom-up statistical analysis that connectionist models appear to employ, even on a scale comparable with that of a child learning his or her first language. In this context, it is also important to note that achieving strong generalization 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 generalization because in most cases they are provided with the lexical categories of words via syntactic tagging, and hence do not actually *learn* this aspect of language. The question of strong generalization 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.

Overall, connectionist models of syntactic processing are at an early stage of development. Current connectionist models of syntax typically use toy fragments of grammar and small vocabularies, and thus have low input representativeness. Nevertheless, these models have good data contact and a reasonable degree of task veridicality. However, more research is required to decide whether promising initial results can be scaled up to deal with the complexities of real language, or whether a purely connectionist approach is beset by fundamental limitations, so that connectionism can only succeed by providing reimplementations of symbolic methods (see the chapters in Part II of this volume for further discussion).

LANGUAGE PRODUCTION

In connectionist psycholinguistics, as in the psychology of language in general, there is relatively little work on language production. However, some important steps have been taken, most notably by Dell and colleagues. Dell's (1986) spreading activation model of retrieval in sentence production constitutes one of the first connectionist attempts to account for speech production.[6] 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 bidirectionally to each other in a straightforward manner without lateral connections within layers, 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 downward in the network. Activation also spreads upward 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 (1986) 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 empirical predictions concerning the retrieval of phonological forms during production, some of which were later confirmed experimentally in Dell (1988).

Dell's (1986) account of speech errors and the phonological encoding of words has had a considerable impact on subsequent models of speech production, both the connectionist (e.g., Harley, 1993) as well as the more symbolic kind (e.g., Levelt, 1989). More recently, Dell, Schwartz, Martin, Saffran, and Gagnon (1997) used an updated version of this model to fit error data from twenty-one aphasics and sixty normal controls. This network has three layers, corresponding to semantic features, words, and phonemes, with the word units connected bidirectionally to the other layers. It maps from semantic features denoting a concept to a choice of word, and then to the phonemes realizing that word. The model distinguishes itself from the interactive activation models, such as TRACE, by incorporating a two-step approach to production. First, activation at the semantic features spreads throughout the network for a fixed time. The most active word unit (typically the best match to the semantic features) is "selected," and its activation boosted. Second, activation again spreads throughout the network for a fixed time, and the most highly activated phonemes are selected, with a phonological frame that specifies the sequential ordering of the phonemes.

Even in normal production, processing sometimes breaks down, leading to semantic errors (*cat* → *dog*), phonological errors (*cat* → *hat*), mixed semantic and phonological errors (*cat* → *rat*), nonword errors (*cat* → *zat*), and

unrelated errors (*cat* → *fog*). Dell, Schwartz, et al. (1997) propose that normal and aphasic errors reflect the same processes, differing only in degree. Therefore, they set their model parameters by fitting data from controls relating to the five types of errors listed. To simulate aphasia, the model was "damaged" by reducing two global parameters (connection weight and decay rate), leading to more errors. Adjusting these parameters, Dell et al. modeled the five types of errors found for twenty-one aphasics, as well as derived and confirmed predictions about the effect of syntactic categories on phonological errors (*dog* → *log*), phonological effects on semantic errors (*cat* → *rat*), naming error patterns after recovery, and errors in word repetition.

Despite their impressive empirical coverage, these spreading activation models nonetheless suffer from a number of shortcomings. As previously mentioned, in 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 if it assumed that the relevant linguistic knowledge is innate, but the information encoded in Dell's (1986) model is language-specific and could not be innate. There is, however, a more urgent, practical side of this problem. It is very difficult to scale these models up, because hand coding becomes prohibitorily complex as the number of weights in the network increases. This shortcoming is alleviated by a recent recurrent network model presented by Dell, Juliano, and Govindjee (1993). The model learns to form mappings from lexical items to the appropriate sequences of phonological segments. The model consists of an SRN with an additional modification: the current output, as well as the current hidden-unit state, are 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. 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, Juliano, et al. (1993) 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; see Dell, Chang, & Griffin, Chapter 7, this volume, for further discussion; but cf. Dell, Burger, & Svec, 1997). It is important that this recent connectionist model suggests that sequential biases and similarity may explain aspects of human phonology that 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.

Connectionist models have also been applied to experimental data on sentence production, particularly concerning structural priming. Structural priming arises when the syntactic structure of a previously heard or spoken

sentence influences the processing or production of a subsequent sentence. Chang, Griffin, Dell, and Bock (1997) (see also Dell et al., Chapter 7, this volume) present an SRN model of grammatical encoding, suggesting that structural priming may be an instance of implicit learning (i.e., acquiring sequential structure with little or no conscious awareness of doing this; see Cleeremans, Destrebecqz, & Boyer, 1998, for a review). This model can be seen as an extension of the Dell, Juliano, et al. (1993) approach. The input to the model was a "proposition," coded by units for semantic features (e.g., *child*), thematic roles (e.g., *agent*) and action descriptions (e.g., *walking*), and some additional input encoding the internal state of an unimplemented comprehension network. The network outputs a sequence of words expressing the proposition. Structural priming was simulated by allowing learning to occur during testing. This created transient biases in the weights of the network, and these are sufficiently robust to cause the network to favor (i.e., to be primed by) recently encountered syntactic structures.

Chang, Griffin, et al. (1997) fitted data from Bock and Griffin (in press) concerning the priming, across up to ten unrelated sentences, of active and passive constructions as well as prepositional (*The boy gave the guitar to the singer*) and double-object (*The boy gave the singer the guitar*) dative constructions. The model fitted the passive data well, and showed priming from intransitive locatives (*The 747 was landing by the control tower*) to passives (*The 747 was landed by the control tower*). However, it fitted the dative data less well, and showed no priming from transitive locatives (*The wealthy woman drove the Mercedes to the church*) to prepositional datives (*The wealthy woman gave the Mercedes to the church*). Chang, Dell, Bock, and Griffin (2000) provide a better fit to these data with a model combining the production network with an implemented comprehension network, and employing a more "fuzzy" representation of thematic roles.

The connectionist production models make good contact with the data, and have reasonable task veridicality, but suffer from low input representativeness, as they are based on small fragments of natural language. It seems likely that connectionist models will continue to play a central role in future research on language production; scaling up these models to deal with more realistic input is a major challenge for future work.

READING

The psychological processes engaged in reading are extremely complex and varied, 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 simple aspects of reading: (1) recognizing letters and words from printed text, and (2) word "naming" (i.e., mapping visually presented letter strings onto sequences of sounds). We focus on models of these two processes here.

One of the earliest connectionist models was McClelland and Rumelhart's (1981) interactive activation model of visual word recognition (see also Rumelhart & McClelland, 1982). This network has three layers of units standing for visual features of letters, whole letters (in particular positions within the word), and words. The model uses the same principles as TRACE, but without the need for a temporal dimension, as the entire word is presented at once.

Word recognition occurs as follows. A visual stimulus is presented, which activates in a probabilistic fashion visual feature units in the first layer. As the features become activated, they send activation via their excitatory and inhibitory connections to the letter units, which, in turn, send activation to the word units. The words compete via their inhibitory connections, and reinforce their component letters via excitatory feedback to the letter level (there is no word-to-letter inhibition). Thus, an "interactive" process occurs: Bottom-up information from the visual input is combined with the top-down information flow from the word units. This process involves a cascade of overlapping and interacting processes: Letter and word recognition do not occur sequentially, but overlap and are mutually constraining.

This model accounted for a variety of phenomena, mainly concerning context effects on letter perception. For example, it captures the fact that letters presented in the context of a word are recognized 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 nonwords (i.e., pseudowords, such as *mave*, which are consistent with English phonotactics) are recognized 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 nonword 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, the model provides an impressively detailed demonstration of how interactive processing can account for a range of further experimental effects.

The interactive activation model of reading is closely related with the TRACE model of speech perception, and explains effects of linguistic context on letter or phoneme perception in a similar way. If the interactive activation framework is appropriate in both domains, then we should expect that the pattern of data in speech and reading should show striking similarities. In line with this expectation, striking parallels between contextual effects in speech perception and reading continue to be discovered. Recently, for example, Jordan, Thomas, and Scott-Brown (1999) demonstrated a "graphemic restoration effect" that parallels the phonemic restoration effect. If a word is viewed from a long distance with some letters deleted and replaced by "noise" (e.g., a spurious character), the "missing" letters are frequently subjectively "seen," just as people report hearing phonemes that have been replaced by noise in the phoneme restoration effect.

But even if strong parallels between speech perception and reading can be established, this connection can, of course, cut both ways. Proponents of a bottom-up view of speech perception can argue that a bottom-up approach can also deal with contextual effects found in reading. Thus, Massaro (1979) has argued that in the context of reading, just as in speech perception, the bottom-up fuzzy logic model of perception provides a better account of the data. Similarly, Norris (e.g., 1994), also a strong advocate of bottom-up models in speech perception, has developed bottom-up accounts of reading. The debate between bottom-up and interactive accounts remains unresolved, although, as we shall see, bottom-up connectionist accounts have been more popular than interactive accounts in the next aspect of reading that we consider: word naming rather than word recognition.

Recent connectionist models of reading have focused not on word recognition but on word naming, which involves relating written word forms to their pronunciations. The first such model was Sejnowski and Rosenberg's (1987) NETtalk, which learns to read aloud from text. NETtalk is a two-layer feed-forward net, with input units representing a "window" of consecutive letters of text and output units representing the network's suggested pronunciation for the middle letter. The network pronounces a written text by shifting the input window across the text, letter by letter, so that the central letter to be pronounced moves onward a letter at a time. In English orthography there is not, of course, a one-to-one mapping between letters and phonemes. NETtalk relies on a rather ad hoc strategy to deal with this: In clusters of letters realized by a single phoneme (e.g., "th," "sh," "ough"), only one letter is chosen to be mapped onto the speech sound, and the others are not mapped onto any speech sound. NETtalk learns from exposure to text associated with the correct pronunciation using back-propagation (Rumelhart et al., 1986). Its pronunciation is good enough to be largely comprehensible when fed through a speech synthesizer.

Sejnowski and Rosenberg gained some insight into what their network was doing by computing the average hidden-unit activation given each of a total of seventy-nine 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 seventy-nine 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 network had learned to treat vowels different from consonants. Moreover, the network had learned a number of subregularities among the letter-to-sound combinations (e.g., evidenced 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 feed-forward 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 bidirectional, 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 2,897 monosyllabic words of English, rather than attempting to deal with unrestricted text like NETtalk. Inputs and outputs used the highly distributed wickelfeature type of representation that proved so controversial in the context of past-tense models, as discussed earlier.

The net's performance captured a wide range of experimental data (on the reasonable assumption that the net's error can be mapped onto 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).

As with the past-tense debate, a controversial claim concerning this reading model was that it uses a single route to handle a quasi-regular mapping. This contrasts with the standard view of reading, which assumes that there are two (nonsemantic) routes in reading, a "phonological route," which applies rules of pronunciation, and a "lexical route," which is simply a list of words and their pronunciations. Regular words can be read using either route, but irregulars must be read by using the lexical route and nonwords must use the phonological route (these will not be known by the lexical route). Seidenberg and McClelland (1989) claim to have shown that this dual-route view is not necessarily correct, because their single route can pronounce both irregular words and nonwords. Moreover, they have provided a fully explicit computational model, while previous dual-route theorists had merely sketched the reading system at the level of "boxes and arrows" (though see Coltheart, Curtis, Atkins, & Haller, 1993, and Coltheart & Rastle, 1994, for recent exceptions).

A number of criticisms have been leveled at Seidenberg and McClelland's account. Besner, Twilley, McCann, and Seergobin (1990) argued that the model's nonword reading is actually very poor compared with people (though see Seidenberg & McClelland, 1990). Moreover, Coltheart et al. (1993) argued that better performance at nonword reading can be achieved by symbolic learning methods, using the same word-set as Seidenberg and McClelland.

As in the past-tense debate, the wickelfeature representation has been criticized, leading to alternative representational schemes. For example, Plaut and McClelland (1993) and Plaut, McClelland, Seidenberg, and Patterson (1996) use a localist code that exploits regularities in English orthography and phonology to avoid a completely position-specific representation. Specifically, Plaut et al. segment monosyllable words into onset, vowel, and coda, and orthographic units can stand for groups of letters (e.g., wh, ea, and so on) that can correspond to a single phoneme. Their model learns to read nonwords very well, but it does so by building in a lot of knowledge into the representation, rather than having the network learn this knowledge. One could plausibly assume (cf. Plaut et al.) that some of 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 idea was explored by Harm, Altmann, and Seidenberg (1994), who showed that pretraining a network on phonology can help learning the mapping from orthography to phonology, and was further developed by Harm and Seidenberg (1999), to which we will return.

One problem with this representational scheme, however, is that it only works for monosyllabic words. Bullinaria (1997), on the other hand, also obtains very high nonword reading performance for words of any length. He gives up the attempt to provide a single-route model of reading and aims to model the phonological route, using a variant of NETtalk in which orthographic and phonological forms are not prealigned by the designer. Instead of having a single output pattern, the network has many output patterns corresponding to all possible alignments between phonology and orthography. All possibilities are considered, and the one that is nearest to the network's actual output is taken as the correct output, and used to adjust the weights. This approach, like NETtalk, uses an input window that 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.

Another limitation of the Seidenberg and McClelland (1989) 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 et al. (1996) demonstrated that a feed-forward network can be trained successfully using the actual frequencies of words instead of their log frequency, even to a level of performance similar to that of human subjects on both word and nonword pronunciation.[7]

Connectionist models of reading have been criticized more generally for not modeling effects of specific lexical items (Spieler & Balota, 1997). One defense is that current models are too partial (e.g., containing no letter recognition and phonological output components) to be expected to model word-level effects (Seidenberg & Plaut, 1998). But Plaut (Chapter 8, this volume) takes up the challenge in relation to word-length effects, and trains an SRN to pronounce words phoneme by phoneme. The network can also refixate on the input when unable to pronounce part of a word. The model performs well on words and nonwords, and provides a reasonably good fit with the empirical data on word-length effects (e.g., Rastle & Coltheart, 1998; Weekes, 1997). These encouraging results suggest that the model may provide a first step toward a connectionist account of the temporal aspects of reading. Complementary work by Harm and Seidenberg (1999) using a recurrent network focuses on providing a richer model of phonological knowledge and processing, which is widely viewed as importantly related to reading and reading development (e.g., Bradley & Bryant, 1983; Goswami & Bryant, 1990).

A further difficulty for Seidenberg and McClelland's (1989) model is the apparent double dissociation between phonological and lexical reading in acquired dyslexia: Surface dyslexics (Bub, Cancelliere, & Kertesz, 1985; McCarthy & Warrington, 1986) can read exception words but not nonwords, but phonological dyslexics (Funnell, 1983) can pronounce nonwords but not irregular words. The standard (although not certain) inference from double dissociation to modularity of function suggests that normal nonword and exception-word reading are subserved by distinct systems, leading to a dual-route model (e.g., Morton & Patterson, 1980). 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 produce this 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 feed-forward networks trained by back-propagation, and concluded that while double dissociations between rules and exceptions can occur in single-route models, this appears to occur only in very small-scale networks. In large networks the dissociation in which the rules are damaged but the exceptions are preserved does not occur. It remains possible that a 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.[8] However, Bullinaria and Chater's results indicate that modeling phonological dyslexia is potentially a major challenge for any single-route connectionist model of reading.

Single- and dual-route theorists argue about whether nonword and exception-word reading is carried out by a single system, but agree that there is an additional "semantic" route, in which pronunciation is retrieved via a semantic code. This pathway is evidenced by deep dyslexics, who make semantic errors in reading aloud, such as reading the word *peach* aloud as "apricot." Interestingly, the behavior of the putative semantic route used by deep dyslexics has itself been modeled using connectionist methods (Hinton & Shallice, 1991; Plaut & Shallice, 1993). Roughly, a back-propagation network is trained to form a highly distributed mapping between words and their meanings (a mapping that is largely, although not completely, arbitrary). If such a network is damaged, then the resulting pattern of errors can involve confusing visually similar or semantically similar words, as well as a surprisingly large number of errors that apparently have both a visual and a semantic component. The profile of errors produced by networks of this kind seems to map, at least in a qualitative way, onto the patterns observed in deep dyslexics. The semantic route used by deep dyslexics is, according to Plaut et al. (1996), also involved in normal reading. In particular, they suggest that a division of labor emerges between the phonological and the semantic pathways during reading acquisition. Roughly, the phonological pathway moves toward a specialization in regular (consistent) orthography-to-phonology mappings at the expense of exception words, which are read by the semantic pathway.

The putative effect of the latter pathway was simulated by Plaut et al. (1996) as extra input to the phoneme units in a feed-forward network trained to map orthography to phonology. The strength of this external input is frequency dependent and gradually increases during learning. As a result, the network comes to rely on this extra input. If eliminated (following a simulated lesion to the semantic pathway), the net loses much of its ability to read exception words, but retains good reading of regular words as well as nonwords. Thus, Plaut et al. provide a more accurate account of surface dyslexia than Patterson et al. (1989). Conversely, selective damage to the phonological pathway (or to phonology itself) should produce a pattern of deficit resembling phonological dyslexia: reasonably good word reading but impaired nonword reading. However, this hypothesis was not tested directly by Plaut et al.

The Plaut et al. (1996) account of surface dyslexia has been challenged by the existence of patients with considerable semantic impairments but who demonstrate a near-normal reading of exception words. Plaut (1997) presents simulations results, suggesting that variations in surface dyslexia

may stem from premorbid individual differences in the division of labor between the phonological and semantic pathways. In particular, if the phonological pathway is highly developed prior to lesioning, a pattern of semantic impairment with good exception-word reading can be observed in the model.

Whereas Seidenberg and McClelland (1989) and Plaut et al. (1996) defend the viewpoint that there is just one nonsemantic route in reading, it is also possible for computational models of reading to embody the opposite view, that there are two nonsemantic routes in reading. Coltheart et al. (1993) have implemented a nonlexical route in which the "grapheme–phoneme correspondences" embodying regular English pronunciation is a nonconnectionist, symbolic algorithm. A second lexical pathway is modeled as a connectionist interactive activation network (Coltheart & Rastle, 1994), building on McClelland and Rumelhart's (1981) model of letter and word recognition, described earlier. This implementation of the dual-route view closely follows previous dual-route theoretical proposals (Baron & Strawson, 1976; Coltheart, 1978; Morton & Patterson, 1980). One route is specifically designed to learn grapheme–phoneme correspondences; the other is specifically designed to read whole words.

In Coltheart's models, the characteristics of each of the two routes are built in, rather than emerging in some natural way from the constraints of the learning task. Zorzi, Houghton, and Butterworth (1998a, 1998b; see also Zorzi, 2000) suggest an elegant alternative. They show that the grapheme–phoneme correspondences in English monosyllabic words can be learned by a connectionist network with no hidden units. Input and output are represented in a simple localist code. There is a separate unit for each letter at each location in the word, and the alignment of the letters and phonemes represents letter positions in relation to the onset–rime structure of the word (the onset is the consonant cluster before the vowel, if any, and the rime is the rest of the word) (Zorzi et al., 1998a). This route learns to read regular words and nonwords correctly, but it cannot deal with exception words effectively. Zorzi et al. (1998b) consider a standard feed-forward network that has a "direct" path from orthography to phonology (i.e., there are no hidden units as before), but which also has an "indirect" path, mediated by a single layer of hidden units. Training this network using standard back-propagation (Rumelhart et al., 1986) leads to an automatic decomposition of the reading task into two functionally separate procedures. Whereas the direct pathway learns grapheme–phoneme correspondences, the indirect pathway uses its hidden units to deal with the word-specific information required to handle exception words. The overall reading performance of this very simple model is surprisingly good, and lesions to the direct or indirect routes produce errors broadly in line with the patterns observed in phonological and surface dyslexia.

Finally, a recent connectionist model has provided insights into developmental rather than acquired dyslexia. Harm and Seidenberg (1999) trained a

network to read in two stages, embodying the observation that children clearly learn a great deal about the phonology of their natural language before learning to map written material onto that phonology. First, they trained a subnetwork consisting of units representing phonetic features to learn the structure of monosyllabic English words. This was done by training the network to auto-associate patterns representing words via a layer of "clean-up" units: The idea is that, after training, the clean-up units are able to correct any errors or omissions in the phonetic representation of a word (the idea of clean-up units had previously been used by Hinton & Shallice, 1991, and Plaut & Shallice, 1993, in the context of cleaning up semantic representations in models of deep dyslexia). To restore errors in the phonetic representation effectively, the subnetwork has to learn the regularities of English phonology. Then they trained a back-propagation network to map orthography to phonology, where the output units that embodied the phonological representation were still associated with the clean-up units. After both phases of training, the resulting network showed a good level of reading performance and replicated the main findings of previous reading models (Seidenberg & McClelland, 1989; Plaut et al., 1996). Moreover, the model shows the potential significance that phonological knowledge may play in assisting reading development: The model trained in two stages learns more quickly than a model that is given no pretraining on phonology.

Harm and Seidenberg (1999) argue that different kinds of damage to the model give rise to analogues of developmental phonological dyslexia and developmental surface dyslexia. This is surprising, because the network does not have separate phonological and lexical reading routes; instead, it is a single homogeneous network. Specifically, Harm and Seidenberg argue that phonological dyslexia can be modeled by impairing the phonological knowledge learned in the first stage of training; for example, by imposing a "decay" on weights in the trained phonological subnetwork or, more drastically, removing the clean-up units entirely. These kinds of damages explain developmental phonological dyslexia in terms of an underlying difficulty with phonological processing, and hence predict that developmental phonological dyslexics will have difficulties with, for example, phonological awareness tasks, as well as difficulties learning the grapheme–phoneme correspondence rules of English. This prediction appears to be born out in the literature (e.g., Share, 1995). By contrast, Harm and Seidenberg view what is often termed developmental surface dyslexia as no more than a delay in the development of normal reading. If this hypothesis is right, then the pattern of reading performance for children with this disorder should be similar to that of younger normal readers. Thus, for example, developmental surface dyslexics should be impaired in their reading not just of irregular words, but also of regular words, when compared to age-matched controls. Harm and Seidenberg argue that this slowing could arise from a number of factors, including lack of exposure to written materials or an inappropriate "learning rate." A learning rate may be inappropriate if it is either too

small, slowing the learning process unnecessarily, or too large, so that the weights jump about excessively rather than converging on a "good" solution. In their simulations, Harm and Seidenberg adopt yet a further approach, slowing learning by using too few hidden units in the back-propagation network mapping orthography to phonology (this approach was previously explored in a preliminary way by Seidenberg & McClelland, 1989). Harm and Seidenberg provide detailed comparisons of the performance of their accounts of both phonological and surface forms of developmental dyslexia with the empirical data. This model therefore stands as a powerful challenge to conventional two-route views of developmental dyslexia (e.g., Castles & Coltheart, 1993; Coltheart et al., 1993). Indeed, Harm and Seidenberg have also shown how a strong double dissociation between reading nonwords and exception words can arise using a single homogeneous network. Although suggestive in relation to similar arguments from neuropsychology, as discussed earlier, it remains to be shown that a similarly crisp pattern of dissociation can be obtained in modeling acquired rather than developmental dyslexia.

Overall, it is clear that the debate between single- and dual-route accounts of nonsemantic reading have not been settled by the growing prevalence of connectionist models. But the advent of connectionist modeling has shifted the debate from typically qualitative discussions of the rival accounts to increasingly sophisticated computational models of the rival positions, which are explicit and produce testable empirical predictions concerning normal reading and the acquired dyslexias. More generally, connectionist models of reading have become central to theory building in the study of reading, and have therefore had a substantial influence on the direction of related experimental and neuropsychological research. With respect to the three criteria for connectionist psycholinguistics, connectionist research on reading has good data contact and reasonable input representativeness. Task veridicality is open to questioning: Children may typically not directly associate written and spoken forms for individual words when learning to read (though Harm and Seidenberg, 1999, partially address this issue). A major challenge for future research is to synthesize the insights gained from detailed models of different aspects of reading into a single model.

PROSPECTS FOR CONNECTIONIST PSYCHOLINGUISTICS

We have seen that controversy surrounds both the past and current significance of connectionist psycholinguistics. Current connectionist models as exemplified in Part I of this volume involve drastic simplifications with respect to real natural language. How can connectionist models be scaled up to provide realistic models of human language processing? Part II provides three different perspectives on how connectionist models may develop.

Seidenberg and MacDonald (Chapter 9, this volume) argue that connectionist models will be able to replace the currently dominant symbolic models of language structure and language processing throughout the cognitive science of language. They suggest that connectionist models exemplify a probabilistic rather than a rigid view of language that requires the foundations of linguistics as well as the cognitive science of language more generally to be radically rethought.

Smolensky (Chapter 10, this volume), by contrast, argues that current connectionist models alone cannot handle the full complexity of linguistic structure and language processing. He suggests that progress requires a match between insights from the generative grammar approach in linguistics and the computational properties of connectionist systems (e.g., constraint satisfaction). He exemplifies this approach with two grammar formalisms inspired by connectionist systems, Harmonic Grammar and Optimality Theory.

Steedman (Chapter 11, this volume) argues that claims that connectionist systems can take over the territory of symbolic views of language, such as syntax or semantics, are premature. He suggests that connectionist and symbolic approaches to language and language processing should be viewed as complementary, but as currently dealing with different aspects of language processing. Nonetheless, Steedman believes that connectionist systems may provide the underlying architecture on which high-level symbolic processing occurs.

Whatever the outcome of these important debates, we note that connectionist psycholinguistics has already had an important influence on the psychology of language. First, connectionist models have raised the level of theoretical debate in many areas by challenging theorists of all viewpoints to provide computationally explicit accounts. This has provided the basis for more informed discussions about processing architecture (e.g., single- versus dual-route mechanisms and interactive versus bottom-up processing). Second, the learning methods used by connectionist models have reinvigorated interest in computational models of language learning (Bates & Elman, 1993). While Chomsky (e.g., 1986) has argued for innate "universal" aspects of language, the vast amount of language-specific information that the child acquires must be learned. Connectionist models may account for how some of this learning occurs. Furthermore, connectionist models provide a test bed for the learnability of linguistic properties previously assumed to be innate. Finally, the dependence of connectionist models on the statistical properties of their input has contributed to the upsurge of interest in statistical factors in language learning and processing (MacWhinney, Leinbach, Taraban, & McDonald, 1989; Redington & Chater, 1998).

Connectionist psycholinguistics has thus already had considerable influence on the psychology of language. But the final extent of this 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 approximated by rules of grammar. Or perhaps connectionist learning methods do not scale up and connectionism can only succeed by reimplementing standard symbolic models. The future of connectionist psycholinguistics 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 READINGS

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

Background

The PDP volumes (McClelland & Rumelhart, 1986, and Rumelhart & McClelland, 1986b) provide a solid introduction to the application of connectionist networks in cognitive models. Smolensky (1988) offers a connectionist alternative to viewing cognition as symbol manipulation, whereas Fodor and Pylyshyn (1988) is a classic critique of connectionism. Elman et al. (1996) details a more recent, broad perspective on connectionism and cognitive development, but see Marcus (1998) for an opposing view. For further discussions, see Seidenberg and MacDonald (Chapter 9, this volume) and Smolensky (Chapter 10, this volume). Finally, McLeod, Plunkett, and Rolls (1998) is a good introduction to the art of conducting connectionist simulations. The book includes simulators for PC and MacIntosh computers as well as exercises with the major network architectures discussed in this chapter.

Speech Processing

The influential TRACE model of speech perception is described in McClelland and Elman (1986). The empirical study of the compensation for coarticulation is found in Elman and McClelland (1988). Gaskell and Marslen-Wilson (Chapter 3, this volume) explore issues related to a connectionist, bottom-up approach to spoken-word recognition. Turning to word segmentation, Cairns et al. (1997) and Christiansen, Allen, et al. (1998) present two models of this area of speech processing.

Morphology

The classic connectionist model of English past-tense acquisition is Rumelhart and McClelland (1986a), with Pinker and Prince (1988) providing the first comprehensive criticism of this model. See Plunkett and Marchman (1993) and Pinker (1991) for recent updates. Over time the debate has also spread to other areas of inflectional morphology, such as the acquisition of English noun plurals (dual mechanism, Marcus, 1995; single mechanism, Plunkett & Joula, Chapter 4, this volume), as well as cross-linguistically to the acquisition of German noun plurals (dual mechanism, Clahsen et al., 1993; single mechanism, Hahn & Nakisa, in press).

Sentence Processing

Elman (1990, 1991) presents 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. Christiansen and Chater (Chapter 5, this volume) extend this approach to cover complex recursive processing, whereas Tabor and Tanenhaus (Chapter 6, this volume) investigate the effects of semantic role expectations on sentence processing. For a structured connectionist approach to the processing of sentences, see Miikkulainen (1996). Steedman (Chapter 11, this volume) provides a critical perspective on connectionist models of syntax.

Language Production

The classic spreading activation model of speech production and speech errors is Dell (1986). Dell et al. (Chapter 7, this volume) describe three subsequent models: an extension to the original model applied to the modeling of aphasic patient data, a bottom-up alternative to the original model of speech errors, and a model of syntactic priming in sentence production. Other connectionist approaches are found in Harley (1993), among others.

Reading

The early interactive activation model of visual word recognition is found in McClelland and Rumelhart (1981). Seidenberg and McClelland (1989) is the classic single-mechanism, connectionist model 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), Harm and Seidenberg (1999), and Zorzi et al. (1998b), as well as Plaut (Chapter 8, this volume).

NOTES

This work was partially supported by the Leverhulme Trust and by European Commission Grant RTN-HPRN-CT-99-00065 to Nick Chater.

1. The term "connectionism," referring to the use of artificial neural networks to model cognition, was coined by Feldman and Ballard (1982).

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

3. Wickelfeatures are generated in a similar way to wickelphones. The latter involve decomposing a phoneme strings into consecutive triples. Thus, the phoneme string /kæt/ (*cat*) is decomposed into the /_kæ/, /kæt/, and /æt_/. 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). Wickelfeatures correspond to triples of phonetic features rather than triples of entire phonemes.

4. 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*). As pointed out by Pinker and Prince (1988), the Rumelhart and McClelland (1986a) model was not able to adequately accommodate the subregularities.

5. In these simulations, novel words corresponded to units that had not been activated during training.

6. 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 (1986) work was not.

7. 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. They also employed a different representational scheme (due to Plaut & McClelland, 1993) than Seidenberg and McClelland (1989).

8. Whereas "regularity" (the focus of the Bullinaria & Chater, 1995, 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).

REFERENCES

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

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language* (pp. 115–151). Mahwah, NJ: Lawrence Erlbaum.

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

Bach, E., Brown, C., & Marslen-Wilson, W. (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes, 1*, 249–262.

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 386–392.

Bates, E. A., Bretherton, I., & Snyder, L. (1988). *From first word to grammar: Individual differences and dissociable mechanisms*. New York: 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: Basic Blackwell.

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

Bock, J. K., & Griffin. Z. M. (in press). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General.*

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

Bradley, L., & Bryant, P. E. (1983). Categorising sounds and learning to read—causal connection. *Nature, 301*, 419–421.

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, 65*, 14–21.

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

Bub, D., Cancelliere, A., & Kertesz, A. (1985). Whole-word and analytic translation of spelling to sound in a non-semantic reader. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 15–34). London: Lawrence Erlbaum.

Bullinaria, J. A. (1997). Modelling reading, spelling and past tense learning with artificial neural networks. *Brain and Language, 59*, 236–266.

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.

Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 289–310). London: UCL Press.

Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111–153.

Castles, A., & Coltheart, M. (1993). Varieties of developmental dyslexia. *Cognition, 47*, 149–180.

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

Chang, F., Dell, G. S., Bock, J. K., & Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research, 29*, 217–229.

Chang, F., Griffin, Z. M., Dell, G. S., & Bock, J. K. (1997). *Modeling structural priming as implicit learning*. Poster presented at the Computational Psycholinguistics Conference, August, University of California, Berkeley, CA.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.
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: 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. H. (1994). *Infinite languages, finite minds: Connectionism, learning and linguistic structure*. Unpublished doctoral dissertation, University of Edinburgh.
Christiansen, M. H. (2000). *Intrinsic constraints on the processing of recursive sentence structure*. Manuscript in preparation.
Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.
Christiansen, M. H., & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language, 9*, 273–287.
Christiansen, M. H., & Chater, N. (1999). Towards a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23*, 157–205.
Christiansen, M. H., Conway, C. M., & Curtin, S. (2000). A connectionist single-mechanism account of rule-like behavior in infancy. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 83–88). Mahwah, NJ: Lawrence Erlbaum.
Christiansen, M. H., & MacDonald, M. C. (2000). *Processing of recursive sentence structure: Testing predictions from a connectionist model*. Manuscript in preparation.
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., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences, 2*, 406–416.
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.
Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151–216). London: Academic Press.
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.
Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1197–1211.
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 Thirteenth Annual Conference of the Cognitive Science Society* (pp. 328–333). Hillsdale, NJ: Lawrence Erlbaum.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14*, 597–650.
Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987) Phoneme identification and the lexicon. *Cognitive Psychology, 19*, 141–177.
Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E.C.T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 113–134). Hillsdale, NJ: Lawrence Erlbaum.
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.
Daugherty, K., MacDonald, M. C., 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 Fifteenth Annual Conference of the Cognitive Science Society* (pp. 383–388). Hillsdale, NJ: Lawrence Erlbaum.
Daugherty, K., & Seidenberg, M. S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 259–264). Hillsdale, NJ: Lawrence Erlbaum.
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., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*, 123–147.
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., 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: 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., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
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.

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

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 Sixteenth Annual Conference of the Cognitive Science Society* (pp. 319–324). Hillsdale, NJ: Lawrence Erlbaum.

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.

Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science, 19*, 407–439.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read*. London: Lawrence Erlbaum.

Gow, D. W., & Gordon, P. C. (1995). Lexical and pre-lexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 344–359.

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.

Hahn, U., & Nakisa, R. C. (in press). German inflection: Single or dual route? *Cognitive Psychology*.

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 Eighth Annual Conference of the Cognitive Science Society* (pp. 106–119). Hillsdale, NJ: Lawrence Erlbaum.

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

Hare, M., Elman, J. L., & Daugherty, K. (1995). Defaultgeneralization 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. W., Altmann, L., & Seidenberg, M. S. (1994). Using connectionist networks to examine the role of prior constraints in human learning. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 392–396). Hillsdale, NJ: Lawrence Erlbaum.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491–528.

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.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74–95.

Hoeffner, J. (1997). *Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Howells, T. (1988). VITAL, a connectionist parser. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

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

Jordan, T. R., Thomas, S. M., & Scott-Brown, K. C. (1999). The illusory-letters phenomenon: An illustration of graphemic restoration in visual word recognition. *Perception, 28*, 1413–1416.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kim, J. J., Pinker, S., Prince, A., & 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.

Kučera, 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.

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

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

Linebarger, M. C., Schwartz, M. F., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13*, 361–392.
MacDonald, M. C., & Christiansen, M. H. (in press). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*.
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 [ʃ]–[s] distinction. *Perception and Psychophysics, 28*, 213–228.
Marcus, G. F. (1995). Children's overregularization of English plurals: A quantitative analysis. *Journal of Child Language, 22*, 447–459.
Marcus, G. F. (1998). Can connectionism save constructivism? *Cognition, 66*, 153–182.
Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology, 29*, 189–256.
Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77–80.
Marr, D. (1982). *Vision*. San Francisco: Freeman.
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.
Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*, 653–675.
Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 595–609.
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.
Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logic model of speech perception. *Cognitive Psychology, 21*, 398–421.
Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23*, 558–614.
McCarthy, R., & Warrington, E. K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex, 22*, 359–380.
McClelland, J. L. (1991). Stochastic interactive processes and the effects of context on perception. *Cognitive Psychology, 23*, 1–44.
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: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models* (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: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models* (pp. 272–325). Cambridge, MA: MIT Press.
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375–407.
McClelland, J. L., & Rumelhart, D. E. (Eds.) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. 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.
McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford: Oxford University Press.
McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 433–443.
McQueen, J. M., Norris, D. G., & Cutler, A. (in press). Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*.
McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 282–312.
Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science, 20*, 47–73.
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. 2, pp. 419–491). New York: Wiley.
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 Fifteenth Annual Conference of the Cognitive Science Society* (pp. 759–764). Hillsdale, NJ: Lawrence Erlbaum.
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* (pp. 91–118). 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: The Second Sperlonga Meeting* (pp. 211–234). Hillsdale, NJ: Lawrence Erlbaum.

Norris, D. G. (1994). A quantitative multiple-levels model of reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1212–1232.

Norris, D. G., McQueen, J. M., & Cutler, A. (in press). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*.

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.

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.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347–370.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 699–725.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes, 12*, 765–805.

Plaut, D. C., & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum.

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

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*, 377–500.

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

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Aquiring verb morphology in children and connectionists. *Cognition, 48*, 21–69.

Pollack, J. B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 33–39). Hillsdale, NJ: Lawrence Erlbaum.

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

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*, 1–56.

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic Bulletin & Review, 5*, 277–282.

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes, 13*, 129–191.

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

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425–469.

Reilly, R. G. (1984). A connectionist model of some aspects of anaphor resolution. In *Proceedings of the Tenth 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 D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 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: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.) (1986b). *Parallel distributed processing: Explorations in the microstructure of cognition*. 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.

Salmela, P., Lehtokangas, M., & Saarinen, J. (1999). Neural network based digit recognition system for voice dialing in noisy environments. *Information Sciences, 121*, 171–199.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*, 28–51.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*, 97–127.

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

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

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., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9*, 234–237.

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: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models* (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 Seventh Annual Conference of the Cognitive Science Society* (pp. 212–221). Hillsdale, NJ: Lawrence Erlbaum.

Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition, 55*, 151–218.

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*, 1–23.

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). Departament de Psicologia Bàsica, Universitat de Barcelona, Spain.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word recognition down to the item level. *Psychological Science, 8*, 411–416.

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

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 Thirteenth Annual Conference of the Cognitive Science Society* (pp. 908–912). Hillsdale, NJ: Lawrence Erlbaum.

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. L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 414–419). Hillsdale, NJ: Lawrence Erlbaum.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword latency. *Quarterly Journal of Experimental Psychology, 50A*, 439–456.

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

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

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation, 2*, 490–501.

Zorzi, M. (2000). Serial processing in reading aloud: No challenge for a parallel model. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 129–136.

Zorzi, M., Houghton, G., & Butterworth, B. (1998a). The development of spelling–sound relationships in a model of phonological reading. *Language and Cognitive Processes, 13*, 337–371.

Zorzi, M., Houghton, G., & Butterworth, B. (1998b). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1131–1161.