To appear in M. Snowling, C. Hulme, and M. Seidenberg (Eds.), *The science of reading: A handbook*, Blackwell: Oxford.

Connectionist Approaches to Reading

David C. Plaut
Departments of Psychology and Computer Science
and the Center for the Neural Basis of Cognition
Carnegie Mellon University

March 2004

Reading is a highly complex task, involving the rapid coordination of visual, phonological, semantic and linguistic processes. Computational models have played a key role in the scientific study of reading. These models allow us to explore the implications of specific hypotheses concerning the representations and processes underlying reading acquisition and performance. A particular form of computational modeling, known as connectionist or neural network modeling, offers the further advantage of being explicit about how such mechanisms might be implemented in the brain.

In connectionist models, cognitive processes take the form of cooperative and competitive interactions among large numbers of simple neuron-like processing units. Typically, each unit has a real-valued activity level, roughly analogous to the firing rate of a neuron. Unit interactions are governed by weighted connections that encode the long-term knowledge of the system and are learned gradually through experience. Units are often organized into layers or groups; the activity of some groups of units encode the input to the system; the resulting activity of other groups of units encodes the system's response to that input. For example, one group might encode the written form (orthography) of a word, another might encode its spoken form (phonology), and a third might encode its meaning (semantics; see Figure 1). The patterns of activity of the remaining groups of units—sometimes termed "hidden" units—constitute learned, internal representations that mediate between inputs and outputs. In this way, the connectionist approach attempts capture the essential computational properties of the vast ensembles of real neuronal elements found in the brain using simulations of smaller networks of more abstract units. By linking neural computation to behavior, the framework enables developmental, cognitive and neurobiological issues to be addressed within a single, integrated formalism.

One very important advantage of connectionist models is that they deal explicitly with learning. Though many of these models have focussed predominantly on simulating aspects of adult, rather than childrens reading, many of the models do explicitly consider the process of learning (e.g., Plaut, McClelland, Seidenberg & Patterson, 1996; Seidenberg & McClelland, 1989). In essence, such models instantiate learning as a process as a slow incremental increase in knowledge, represented by increasingly strong and accurate connections between different units (e.g., the letters in printed words and the phonemes in spoken words to which they correspond). Another critical feature of many connectionist systems is that after learning they show the ability to generalize (e.g., to pronounce novel words which they have not been trained on). Finally, and related to this, such systems often show graceful degradation when damaged. Removing units or connections in such systems typically does not result in an all-or-none loss of knowledge; rather,

The preparation of this chapter was supported by NIH grant MH55628. Correspondence may be sent to David Plaut (plaut@cmu.edu), Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213–3890.

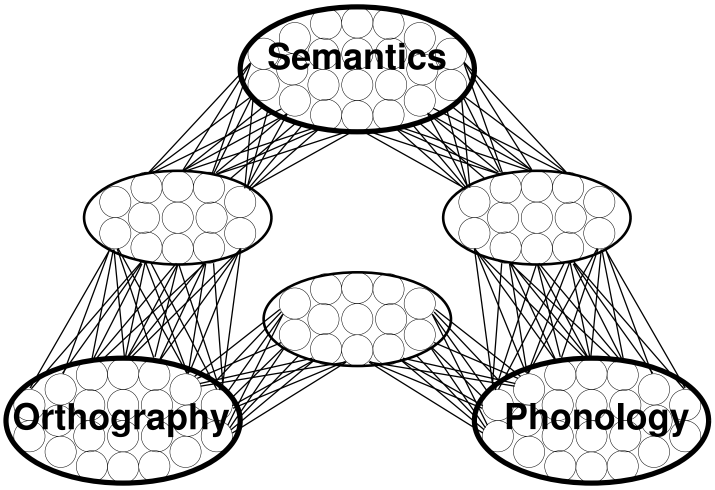

Figure 1. A connectionist network that relates orthographic, phonological, and semantic information in word reading and other lexical tasks, based on the “triangle” framework (Harm & Seidenberg, in press; Plaut, McClelland, Seidenberg & Patterson, 1996; Seidenberg & McClelland, 1989).

damage results in a gradual degradation of performance. These three aspects of connectionist models have clear parallels in human reading behaviour—children gradually learn to read more and more words in an incremental fashion over a long period, such learning brings with it the ability to generalize to novel items children have not been taught, and in cases of brain damage there are often graded declines in performance with inconsistent performance at different times. The fact that connectionist models display such parallels to human reading behaviour has generated considerable excitement at the prospect that such models may offer new, explicit and detailed accounts of how reading is implemented in the human brain.

Principles of Connectionist Modeling

Before turning to how specific connectionist models have been applied to various reading-related phenomena, it will be helpful to consider the implications of the underlying computational principles more generally. These can be grouped into issues related to processing, representation, learning, and network architecture.

Processing

A standard connectionist unit integrates information from other units by first computing its *net input*, equal to a linear sum of positive- and negative-weighted activations from sending units, and then setting its own activation according to a nonlinear, monotonically increasing (sigmoid) function of this net input (see Figure 2). In some networks, unit activations change gradually in response to input from other units instead of being recomputed from scratch each time.

Both the linear integration of net input and the nonlinear activation function play critical roles in shaping how connectionist networks behave. The fact that the net input to each unit is a simple weighted sum is at the heart of why networks exhibit similarity-based generalization to novel inputs (e.g., being able to pronounce a pseudoword like MAVE based on knowledge of words like GAVE, SAVE, MATE, etc.). If a unit is presented with a similar pattern of activity along its input lines, it will tend to produce a similar net input and, hence, a similar response. This fails to hold only if the weights for those inputs that differ between the patterns are very large, but such large weights develop during learning only when necessary (e.g., when handling exceptional cases; see the section on Learning below).

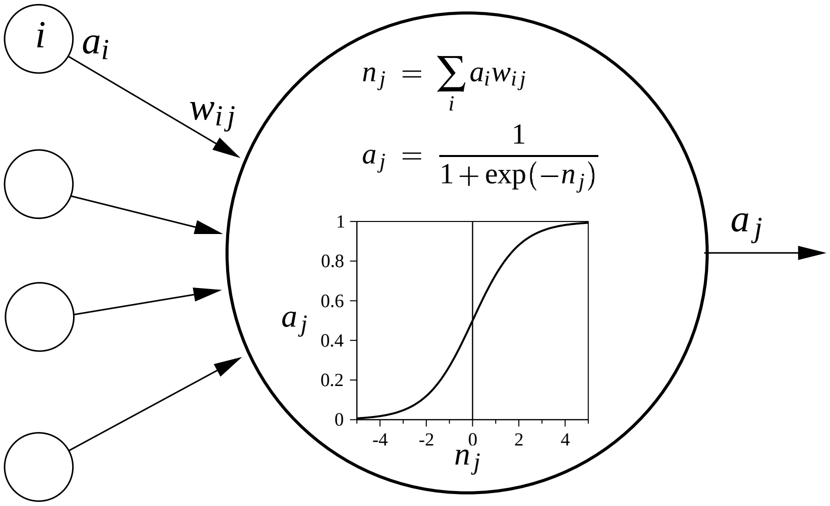

Figure 2. The operation of a standard connectionist unit (indexed by j), which computes a net input n_j as a weighted sum of activations a_i from other units (indexed by i), and then computes its own activation a_j as a smooth, nonlinear (sigmoid) function of its net input (where exp(·) is the exponential function).

If all processing in the network were strictly linear, however, the types of mappings it could learn would be severely limited (Minsky & Papert, 1969). The nonlinear activation function allows individual units—and hence the network as a whole—to preserve some types of similarity in its response while ignoring others. The sigmoid activation function asymptotes for large positive or negative net inputs, but produces roughly proportional responses for small and moderate net inputs (see Figure 2). If networks start out with relatively small weights, most units activations will fall in the linear range of the sigmoid function, and the network as a whole will give similar responses to similar inputs. However, when aspects of a task require responses that are not predicted by input similarity (e.g., pronouncing SEW like SO instead of SUE, or mapping CAP and CAT to completely different meanings), learning must develop sufficiently large weights to drive the relevant units into their nonlinear (asymptotic) range, where changes in net input have little if any effect on activation. In this way, a network can remain largely linear for systematic or “regular” aspects of a task, while simultaneously exhibiting nonlinear behavior for the unsystematic or “irregular” aspects.

Understanding how a connectionist network operates above the level of individual units requires consideration of how patterns of activity across the various groups of units interact and evolve over the course of processing a given input. A very useful concept in this regard is the notion of an *attractor*. At any given instant, the current pattern of activity over a group of units in the network (or over the network as a whole) can be represented in terms of the coordinates of a point in a multi-dimensional *state space* that has a dimension for each unit. As the pattern of activity changes during processing, the corresponding point in state space moves. In many networks, unit interactions eventually reach a state in which the activation of each unit is maximally consistent with those of other units and the pattern as a whole stops changing. The point in state space corresponding to this final pattern is called an attractor because interactions among units in the network cause nearby points (i.e., similar patterns) to be “pulled” towards the same final attractor point. (The region around an attractor that settles to it is called its *basin* of attraction.) The stability of attractor patterns gives networks a considerable degree of robustness to partially missing or noisy input or to the effects of damage.

Representation

As described thus far, a typical connectionist network processes an input through unit interactions that cause the network to settle to an attractor, in which the resulting pattern of activity over output units corresponds to the network's response to the input. An issue of central relevance is the nature of the representations that participate in this process—the way that inputs, outputs, and groups of intermediate units encode information in terms of patterns of activity. Some connectionist models use *localist* representations, in which individual units stand for familiar entities such as letters, words, concepts, and propositions. Others use *distributed* representations, in which each such entity is represented by a particular pattern of activity over many units rather than by the activity of an single unit. Localist representations can be easier to think about and to manipulate directly (Page, 2000), but often permit too much flexibility to constrain theorizing sufficiently (Plaut & McClelland, 2000). By contrast, distributed representations are typically much more difficult to use and understand but can give rise to unanticipated emergent properties that contribute in important ways to the explanation of cognitive phenomena (see, e.g., Hinton & Shallice, 1991).

Given that, as explained above, similar patterns tend to have similar consequences in connectionist networks, the key to the use of distributed representations is to assign patterns to entities in such a way that the similarity relations among patterns captures the underlying functional relationships among the entities they represent. For groups of units that must be interpreted directly (i.e., inputs and outputs), this is done based on independent empirical evidence concerning the relevant representational similarities. However, except for the simplest of tasks, it is impossible to perform the relevant mappings without additional intermediate units, and it is infeasible to specify appropriate connection weights for such units by hand. Accordingly, distributed connectionist networks almost invariably use learning to discover effective internal representations based on task demands.

Learning

The knowledge in a network consists of the entire set of weights on connections among units, because these weights govern how units interact and hence how the network responds to any given input. Accordingly, learning involves adjusting the weights in a way that generally benefits performance on one or more tasks (i.e., mapping from inputs to outputs).

Connectionist learning procedures fall into three broad classes based on how much performance feedback is available. At one extreme are *unsupervised* procedures, such as Hebbian learning (as it is typically applied; Hebb, 1949), that make no use of performance feedback and, instead, adjust connection weights to capture the statistical structure among activity patterns. At the other extreme are *supervised* procedures, such as back-propagation (Rumelhart, Hinton, & Williams, 1986), that assume the learning environment provides, for every trained input pattern, a fully specified "target" pattern that should be generated over the output units. Between these two extremes are *reinforcement* procedures, such as temporal difference methods (Sutton, 1988), that assume the environment provides potentially intermittent evaluative feedback that does not specify correct behavior but rather conveys the degree to which behavioral outcomes were good or bad.

When performance or evaluative feedback is available, it is relatively straightforward to use it to adapt connection weights to improve performance. If the activation of an output unit is too high, it can be reduced by decreasing positive incoming weights and the corresponding sending activations and by increasing (in magnitude) negative weights and sending activations (see the equations in Figure 2); the reverse is true if output activation is too low. Changing the sending activations involves reapplying the same procedure to their incoming weights and incoming activations, and so on. Specific algorithms differ in how they compute feedback and how they distribute information on how to change weights.

Many applications of distributed connectionist modeling to cognitive phenomena use back-propagation despite its biological implausibility (Crick, 1989). This is partly because, unlike most alternatives, the procedure is effective at learning difficult mappings, including those with complex temporal characteristics (Williams & Peng, 1990). It is also the case that the time-course and ultimate outcome of learning with back-propagation is highly similar to the properties of more biologically plausible supervised procedures, such as Contrastive Hebbian learning (Ackley, Hinton, & Sejnowski, 1985; O'Reilly, 1996; Peterson & Anderson, 1987). Thus, one can interpret back-propagation as a computationally efficient means of learning internal representations in distributed connectionist networks in a way that approximates the properties of performance-driven learning in the brain.

Network Architecture

The *architecture* of a network—the pattern of connectivity among and within groups of units representing different types of information—can have an important impact on the behavior of a connectionist model in its acquisition, skilled performance, and impairment following damage. The strong emphasis on learning in the development of connectionist models has led some researchers to conclude that the approach disavows any built-in structure within the cognitive system. A more accurate characterization would be that the effectiveness of learning in connectionist network makes it possible to explore the degree to which built-in structure is necessary to account for some empirical phenomena. The modeling framework itself allows for the expression of a wide variety of network architectures, ranging from those with extensive built-in structure to those with minimal structure.

Connectionist models often contrast with alternative formulations in terms of the *kinds* of distinctions that are instantiated in the architecture of the system. A classic example is the traditional separation of rule-based and item-based mechanisms in "dual-route" theories of word reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and inflectional morphology (Pinker, 1999). Because the processing mechanisms within a connectionist system are homogeneous—involving massively parallel unit interactions throughout—the underlying theories rarely isolate different types of *processing* into separate systems or pathways. Rather, architectural divisions typically reflect different types of *information* (e.g., orthographic, phonological, semantic). Given that such distinctions often correspond to modalities of input or output, they can be supported directly by data on neuroanatomic localization of the corresponding neural representations.

Realist Versus Fundamentalist Approaches

Before turning to an overview of connectionist models of reading, it is worth distinguishing two broad approaches to cognitive modeling, because they often have rather different goals. The *realist* approach tries to incorporate into a model as much detail as possible of what is known about the real system in the belief that complex interactions of these factors are necessary to capture the relevant phenomena. The *fundamentalist* approach, by contrast, holds that a model should, as much as possible, embody only those principles that are claimed to account for the relevant phenomenon and should abstract out extraneous details. In evaluating any given modeling effort, it is important to identify the specific goals of the work; some models are intended to provide comprehensive accounts of detailed behavioral data, whereas others are intended more as demonstrations of specific computational arguments. Often the most effective modeling approach over the long term is to begin with fundamentalist models to elucidate the key underlying principles, and then gradually move towards more realist models as the theoretical implications of additional details become understood.

Connectionist Modeling of Reading

Most connectionist models of reading have focused on single word processing as it is generally thought that, above the lexical level, written language engages largely the same mechanisms as spoken language. In the review that follows, these models are characterized in terms of whether their representions for words are localist (one unit per word) or distributed (alternative patterns of activity for each word) and whether they focus on the task of word recognition (deriving a lexical or semantic representation) or oral reading (deriving a pronunciation).

Localist Models of Word Recognition

One of the earliest and arguably most influential connectionist models of reading is a localist, non-learning model—the Interactive Activation (IA) model of letter and word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). The model consists of three layers of units—letter feature units, letter units, and word units. The model was designed to recognize four-letter words, so there is a separate set of feature units and letter units for each of four letter positions. The activation of each unit can be thought of as reflecting the network's confidence in the hypothesis that the entity represented by the unit (e.g., a T in the first position, or the word TAKE) is part of the correct interpretation. The weights on connections between units reflect the degree to which one hypothesis is consistent or inconsistent with another. Within each level, units representing inconsistent hypotheses (e.g., a T versus a P in the first letter position, or the words TAKE and TRIP) have negative connections between them. Between levels, units representing consistent hypotheses (e.g., a top horizontal letter feature and the letter T, or a T in the first position and the word TAKE) have positive connections between them, whereas units representing inconsistent hypotheses (e.g., a P in the first position and the word TAKE) have negative connections between them. Connections throughout the system are bidirectional, allowing both top-down and bottom-up information to influence unit activations.

A primary goal of the model was to explain the *word superiority effect* (Reicher, 1969; Wheeler, 1970), in which the perception of a briefly presented letter is more accurate when it occurs in a word compared with when it occurs in a random consonant string or even in isolation (see Lupker, this volume). In the IA model, this effect arises due to partial activation of word units that provide top-down support for the letters they contain. The model was also able to explain the *pseudoword superiority effect* (e.g., Carr, Davidson, & Hawkins, 1978; McClelland & Johnston, 1977), in which letters occurring in pronounceable nonwords (e.g., MAVE) are perceived better than in consonant strings or in isolation (although not quite as well as in words). Although pseudowords are not fully consistent with any of the units at the word level in the model, they are partially consistent with many words. The presentation of a pseudoword typically generates weak activation of word units sharing three of its four letters; these units, in turn, conspire to provide top-down support for the letters in the pseudowords. In this way, the IA model provided an early demonstration of how even a localist model can generalize on the basis of similarity, through the use of what are essentially distributed representations for pseudowords.

In subsequent work, McClelland (1991; see also Movellan & McClelland, 2001) elaborated the model to use units with an intrinsically noisy or stochastic activation function to bring the model in line with empirical evidence for statistical independence in how people integrate multiple sources of information (Massaro, 1988). More recently, Grainger and Jacobs (1996) generalized the interactive activation framework to address a broader range of tasks and issues related to word recognition.

Distributed Models of Word Recognition

Mozer (1991) developed a connectionist model of object recognition and spatial attention, called MORSEL, that was applied to the specific task of recognizing words. In the model, an attentional system forms a spatially contiguous bubble of activation that serves to select a subset of the bottom-up letter feature information for further processing by a hierarchically organized object recognition system. Each layer in the recognition system (called BLIRNET) consists of units with spatially restricted receptive fields that form conjunctions of the simpler features in the previous layer. At the top of the system are position-independent units that respond to specific triples of letters (following Wickelgren's, 1969, proposal for representing spoken words). In this way, words were represented by a pattern of activity over multiple letter triples (e.g., #HO, OUS, USE, SE#, for the word HOUSE) rather than by the activation of a single word unit (as in the IA model). Although there was no learning in the system, it was still successful at activating the correct set of letter triples for a fairly large vocabulary of words. When presented with multiple words, it usually selected and recognized one of them accurately but, like human subjects, would occasionally misrecognize the attended word due to letter migrations from the unattended word (Mozer, 1983). Moreover, when one side of the attentional mechanism was impaired, the damaged model exhibited all of the major characteristics of neglect dyslexia, the manifestation of hemispatial neglect with written words as stimuli (Mozer & Behrmann, 1990).

Although MORSEL used distributed word representations, it did not employ learning. Other distributed models have cast the problem of word recognition as mapping from the written forms of words to their meanings (rather than to higher-order orthographic representations, as in MORSEL), and have used learning to develop weights that accomplish this mapping. Note, however, that, apart from morphological relationships, the relationship between the surface forms of words and their meanings is largely arbitrary. In other words, similarity in form (e.g., CAT, CAP) is unrelated to similarity in meaning (e.g., CAT, DOG). This is the most difficult type of mapping for connectionist networks to learn, given their inherent bias towards preserving similarity. In fact, some researchers questioned whether it was even possible for distributed networks to accomplish this mapping without word-specific intermediate units. Kawamoto (1988) used a variant of Hebbian learning to train a distributed network to map among orthographic, phonological and semantic representations (see also Van Orden, Pennington, & Stone, 1990). However, because the network lacked any hidden units, it could learn a vocabulary of only a few words. Nonetheless, Kawamoto was able to show that the model provided a natural account of a number of phenomena related to lexical semantic ambiguity resolution (see also Kawamoto, 1993; Kawamoto, Kello, & Jones, 1994).

To address the more general challenge, Hinton and Sejnowski (1986) trained a Boltzmann Machine—a network of stochastic binary units—to map between orthography and semantics for a larger (although still small) set of words. Although training was difficult, the network was able to develop distributed representations over intermediate hidden units that accomplished the mapping. They also found that, with mild damage, the network occasionally responded to a word by giving another, semantically related word as a response (e.g., CAT read as DOG)—a *semantic error* reminiscent of those made by patients with *deep dyslexia* (Coltheart, Patterson, & Marshall, 1980).

Following Hinton and Sejnowski (1986), Hinton and Shallice (1991) used back-propagation to train a recurrent network with hidden units to map from orthography to semantics for 40 words falling into five concrete semantic categories. Orthographic representations were based on position-specific letter units; semantic representations consisted of subsets of 68 hand-specified semantic features that captured a variety of conceptual distinctions among word meanings. When the network was damaged by removing some units or connections, it no longer settled normally; the initial semantic activity caused by an input would occasionally fall within a neighboring attractor basin, giving rise to an error response. These errors were often semantically related to the stimulus because words with similar meanings correspond to nearby attractors in semantic space. Like deep dyslexic patients, the damaged network also produced errors with visual similarity to the stimulus (e.g., BOG read as DOG) and with both visual and semantic similarity (e.g., CAT read as RAT), due to its inherent bias towards similarity: visually similar words tend to produce similar initial semantic patterns, which can lead to a visual error if the basins are distorted by damage (see Figure 3).

Plaut and Shallice (1993) extended these initial findings in a number of ways. They established the generality of the co-occurrence of error types across a wide range of simulations, showing that it does not depend on specific characteristics of the network architecture, the learning procedure, or the way responses are generated from semantic activity. They also showed that distributed attractor networks exhibited a number of other characteristics of deep dyslexia not considered by Hinton and Shallice (1991), including the occurrence of visual-then-semantic errors, greater confidence in visual as compared with semantic errors, and relatively preserved lexical decision with impaired naming. They also extended the approach to address effects of concreteness on word reading in deep dyslexia. They trained a network to pronounce a new set of words consisting of both concrete and abstract words. Concrete words were assigned far more semantic features than were abstract words, under the assumption that the semantic representations of concrete words are less dependent on the contexts in which they occur (Saffran, Bogyo, Schwartz, & Marin, 1980). As

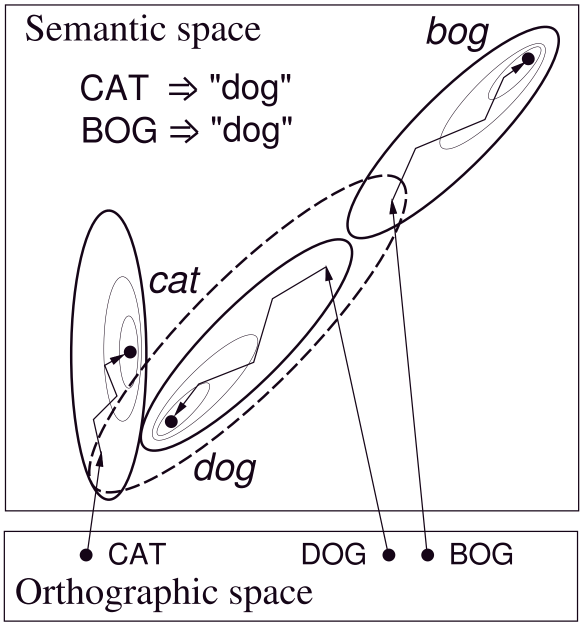

Figure 3. How damage to an attractor network can give rise to both semantic and visual errors. Points within each rectangular area correspond to specific patterns of activation over orthographic or semantic representations; neighboring points corresponding to similar (overlapping) patterns. The arrows reflect the way in which these patterns change over the course of processing. The solid ovals represent the basins of attraction in the normal network; the dashed ovals represent alterations of these basins due to damage. (Based on Hinton & Shallice, 1991)

a result, the network developed stronger attractors for concrete than abstract words during training, giving rise to better performance in reading concrete words under most types of damage, as observed in deep dyslexia. Surprisingly, severe damage to connections implementing the attractors at the semantic level produced the opposite pattern, in which the network read *abstract* words better than concrete words. This pattern of performance is reminiscent of CAV, the single, enigmatic patient with *concrete word dyslexia* (Warrington, 1981). The double dissociation between reading concrete versus abstract words in patients is often interpreted as implying that there are separate modules within the cognitive system for concrete and abstract words. The Plaut and Shallice simulation demonstrates that such a radical interpretation is unnecessary: the double dissociation can arise from damage to different parts of a distributed network which processes both types of items but develops somewhat different functional specializations through learning (see also Plaut, 1995).

"Dual-Route" Models of Reading Aloud

Much of the controversy surrounding theories of word reading centers not around how words are recognized and understood but how they are read aloud. In part, this is because, in contrast to the arbitrary nature of form-meaning mappings, the mapping between the written and spoken forms of words is highly systematic; words that are spelled similarly are typically also pronounced similarly. This property derives from the fact that written English follows an *alphabetic principle* in which parts of written forms (letters and multiletter graphemes like TH, PH) correspond to parts of spoken forms (phonemes). The sharp contrast between the systematic nature of pronunciation and the arbitrary nature of comprehension has led a number of researchers (e.g., Coltheart, 1978; Marshall & Newcombe, 1973) to propose separate pathways or "routes" for these two tasks, each employing very different computational mechanisms: a *sublexical* pathway employing grapheme-phoneme correspondence (GPC) rules for pronunciation, and a *lexical* pathway involving a word-specific lexical look-up procedure for comprehension (characterized much like the IA model in later formulations; see, e.g., Coltheart et al., 2001; Coltheart, this volume). Complications arise, however, because the pronunciation task itself is not fully systematic; roughly 20% of English words are *irregular* in that they violate the GPC rules (e.g., SEW, PINT, YACHT). So-called "dual-route" theories propose that pronouncing such words also depends on the lexical pathway.

Although traditional dual-route models implement the sublexical pathway with symbolic rules (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart et al., 2001), it is perfectly feasible to build a dual-route mechanism out of connectionist hardware. For example, Zorzi, Houghton, and Butterworth (1998) describe simulations in which direct connections from letter units to phoneme units support the pronunciation of regular words and nonwords, whereas a separate pathway, composed either of hidden units or localist word units, supports the pronunciation of irregular words (see also Ans, Carbonnel, & Valdois, 1998). Although the mechanisms employed for the two pathways are more homogeneous than in more traditional, rule-based implementations, the models nonetheless retain a categorical distinction between words that obey spelling-sound rules and words that violate them.

Distributed Models of Reading Aloud

The first researchers to take on the challenge of training a single connectionist network to pronounce all English words were Sejnowski and Rosenberg (1987), who developed a system called NETtalk. Orthographic input was presented to NETtalk by sweeping a 7-letter window over a large text corpus (the Brown corpus; Kucera & Francis, 1967), successively centering the window on each letter in the text. For each letter position, the system was trained to generate the single phoneme corresponding to the central letter in the window. This allows each successive letter to be processed by the same set of units, so the knowledge extracted in processing letters in any position are available for processing letters in every other position. At the same time, the presence of other letters in the surrounding slots allows the network to be sensitive to the context in which letters occur. This is necessary not only for pronouncing exception words but also for handling multiletter graphemes (e.g., TH, PH, SH). For these, the system was trained to generate the appropriate phoneme for the first letter and then silence for the remaining letters. The alignment of phonemes to letters was specified by hand.

Although impressive as a first attempt, the performance of NETtalk when judged in terms of entire words pronounced correctly was much poorer than skilled readers. In follow-up work, Bullinaria (1997) showed that performance in a NETtalk-like system could be improved dramatically by allowing the network to discover the best letter-phoneme alignment by itself. This was done by evaluating the network's output against all possible alignments, and training towards the one that yields the lowest overall error. This pressures the system to converge on alignments that are maximally consistent across the entire training corpus, yielding perfect performance on words and good generalization to pronounceable nonwords.

The need for strictly sequential processing on even the shortest words raises questions about the psychological plausibility of the NETtalk approach. One way to address this concern is to propose that skilled readers attempt to process as much of the input as they can in parallel, then redirect fixation and continue. In this view, unskilled reading may be strictly sequential, as in NETtalk, but as skill develops, it becomes much more parallel. To explore this possibility, Plaut (1999) trained a simple recurrent (sequential) network to produce sequences of single phonemes as output when given position-specific letters as input. The network was also trained to maintain a representation of its current position within the input string. When the network found a peripheral portion of the input difficult to pronounce, it used the position signal to refixate the input, shifting the peripheral portion to the point of fixation where the network had had more experience in generating pronunciations. In this way, the network could apply the knowledge tied to the units at the point of fixation to any difficult portion of the input. Early on in training, the network required multiple fixations to read words, but as the network became more competent it eventually read most words in a single fixation. The network could also read nonwords about as well as skilled readers, occasionally falling back on a refixation strategy for difficult nonwords. Finally, a peripheral impairment to the model reproduced the major characteristics of letter-by-letter reading in pure alexic patients (Behrmann, Plaut, & Nelson, 1998). Specifically, when input letter activations were corrupted with noise, the model exhibited a clear effect of orthographic length in its number of fixations (a loose analog to naming latency), and this effect interacted with lexical frequency such that the increase was much greater for low- compared with high-frequency words.

An alternative approach to word reading, first articulated by Seidenberg and McClelland (1989), casts the problem as learning to map among orthographic, phonological and semantic representations for entire words in parallel (see Figure 1). The approach does not deny the existence of sequential processes related to both visual input and articulatory output, but emphasizes the parallel interactions among more central types of lexical information. In support of this general "triangle" framework, Seidenberg and McClelland (1989) trained a connectionist network to map from the orthography of about 3000 monosyllabic English words—both regular and exception—to their phonology via a set of hidden units (i.e., the bottom portion of the framework in Figure 1, referred to as the *phonological* pathway). The network was also trained to use the same internal representation to regenerate the orthographic input, providing a means for the network of distinguishing words from nonwords based on the accuracy of this reconstruction. Orthographic input was coded in terms of context-sensitive letter triples, much like the highest-level representations in MORSEL. Phonological output was coded in terms of triples of phonemic features. To determine the network's pronunciation of a given letter string, an external procedure constructed the most likely phoneme string given the feature triples generated by the network. This string was then compared with the actual pronunciation of the stimulus to determine whether the network made a correct or error response. After training, the network pronounced correctly 97.7% of the words, including most exception words. The network also exhibited the standard empirical pattern of an interaction of frequency and consistency in naming latency (Andrews, 1982; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985) if its real-valued accuracy in generating a response is taken as a proxy for response time (under the assumption that an imprecise phonological representation would be less effective at driving an articulatory system). However, the model was much worse than skilled readers at pronouncing orthographically legal nonwords and at lexical decision under some conditions (Besner, Twilley, McCann, & Seergobin, 1990). Thus, although highly successful in many respects, the model failed to refute traditional claims that localist, word-specific representations and separate mechanisms are necessary to account for skilled reading.

Plaut, McClelland, Seidenberg, and Patterson (1996) showed, however, that the limitations of the Seidenberg and McClelland model stem not from any general limitation in the abilities of connectionist networks, but from its use of poorly structured orthographic and phonological representations. The triples-based orthographic and phonological representations used by the original model fail to capture the relevant similarities among written and spoken forms of words adequately, essentially because the contribution that each grapheme and phoneme makes is overly sensitive to the surrounding context. When more appropriately structured representations are used—based on graphemes and phonemes and embodying phonotactic and graphotactic constraints—network implementations of the phonological pathway can learn to pronounce regular words, exception words, and nonwords as well as skilled readers. Furthermore, the networks also exhibit the empirical frequency-by-consistency interaction pattern, even when naming latencies are modeled directly by the settling time of a recurrent, attractor network.

Although Plaut et al. (1996) demonstrated that implementations of the phonological pathway on its own can learn to pronounce words and nonwords as well as skilled readers, a central aspect of their general theory is that skilled reading more typically requires the combined support of both the semantic and phonological pathways (see also Hillis & Caramazza, 1991; Van Orden & Goldinger, 1994), and that individuals may differ in the relative competence of each pathway (Plaut, 1997; Seidenberg, 1992). The division-of-labor between these pathways has important implications for understanding acquired surface dyslexia, a neuropsychological disorder in which patients pronounce regular words and nonwords normally but "regularize" exception words, particularly those of low frequency (e.g., SEW read as SUE; see Patterson, Coltheart, & Marshall, 1985). Plaut et al. (1996) explored the possibility that surface dyslexia might reflect the natural limitations of an intact phonological pathway that had learned to rely on semantic support that was reduced or eliminated by brain damage. They approximated the contribution that the semantic pathway would make to oral reading by providing phonological representations with external input that pushed them toward the correct pronunciation of each word during training. A semantic impairment was modeled by weakening this external input. Plaut and colleagues found that, indeed, a phonological pathway trained in the context of support from semantics exhibited the central phenomena of surface dyslexia following semantic damage: intact nonword reading and regularization of low-frequency exception words (see Lambon Ralph & Patterson, this volume). Moreover, as explored in additional simulations (Plaut, 1997), individual differences in the severity of surface dyslexia can arise, not only from differences in the amount of semantic damage, but also from *premorbid* differences in the division of labor between the semantic and phonological pathways. The relative strengths of these pathways, and the overall competence of the reading system, would be expected to be influenced by a wide variety of factors, including the nature of reading instruction, the sophistication of preliterate phonological representations, relative experience in reading aloud versus silently, the computational resources (e.g., numbers of units and connections) devoted to each pathway, and the reader's more general skill levels in visual pattern recognition and in spoken word comprehension and production. On this view, the more severe surface dyslexic patients had greater premorbid reliance on the semantic pathway as a result of one or more of these factors.

A remaining limitation of the Seidenberg and McClelland model that was not addressed by Plaut et al. (1996) concerns the ability of a distributed network lacking word-specific representations to perform lexical decision accurately. The focus of work with the Seidenberg and McClelland model was on demonstrating that, under some conditions, lexical decisions can be performed on the basis of a measure of orthographic familiarity. Plaut (1997) demonstrated that lexical decisions can be made more accurately when based on a familiarity measure applied to semantics. A feedforward network was trained to map from the orthographic representations of the 2998 monosyllabic words in the Plaut et al. (1996) corpus to their phonological representations and to artificially created semantic representations generated to cluster around prototype patterns over 200 semantic features. After training, the network was tested for its ability to perform lexical decision based on semantic *stress*—an information-theoretic measure of the degree to which the states of semantic units differed from rest. When tested on the pronounceable nonwords from Seidenberg, Plaut, Petersen, McClelland, and McRae (1994), there was very little overlap between the semantic stress values for nonwords and those for words: an optimal decision criterion yielded only 1% errors. Moreover, the distributions of stress values for words varied systematically as a function of their frequency. In a second test, the network produced reliably higher semantic stress values—and thus poorer discrimination from words—for the Seidenberg, Petersen, MacDonald, and Plaut (1996) pseudohomophones compared with their controls. Thus, the network exhibited accurate lexical decision performance overall, along with an advantage for higher-frequency words and a disadvantage for pseudohomophones, as found in empirical studies.

More recently, Harm and Seidenberg (in press) have developed a full implementation of the "triangle" framework (see Figure 1) and used it to examine a number of issues related to the division-of-labor in the reading system. Although the focus of the work is on the comprehension of written words via the direct versus phonologically mediated pathways, the underlying principles apply equally well to the computation of phonology both directly or via semantics. First, to approximate preliterate language experience, the network was trained to map bidirectionally between phonology and semantics for 6103 monosyllabic words (see also Harm & Seidenberg, 1999, for a computational examination of the relevance of preliterate experience to reading acquistion). The phonology of each word was encoded in terms of eight slots of 25 phonetic features, organized into a CCCVVCCC template. In constructing semantic representations, words were first categorized by their most frequenct word class (Francis & Kučera, 1982). For uninflected nouns and verbs, semantic features were generated using the WordNet online semantic database (Miller, 1990). Adjectives, adverbs and closed-class words were hand-coded according pre-existing feature taxonomies (e.g., Frawley, 1992). Inflected words were assigned the features of their base forms plus specific inflectional features. In total, 1989 semantic features were generated to encode word meanings, with words averaging 7.6 features each (range 1–37). Once the preliterate network was reasonably accurate at understanding and producing spoken words (86% and 90% correct, respectively), the network was then trained on the reading task. Orthography was encoded using letter units organized into vowel-centered slot-based representation (analogous to phonology). After extended training, the model succeeded in activating the correct semantic features for 97.3% of the words and the correct phonological features for 99.2% of the words.

The trained model exhibited the appropriate effects of word frequency, spelling-sound consistency, and imageability in pronouncing words, and was as accurate as skilled readers in pronouncing pseudowords. Harm and Seidenberg's (in press) primary goal, however, was to address the longstanding debate on whether reading is necessarily phonologically mediated. An examination of the division-of-labor in activating meaning from print over the course of training indicated that the network relied heavily on phonological mediation (orthography-phonology-semantics) in the early stages of reading acquisition but gradually shifted towards increased reliance on the direct mapping (orthography-semantics) as reading skill improved. Even at the end of training, however, both pathways continue to make important contributions to performance. This is especially true for homophones (e.g., ATE, EIGHT), which cannot be comprehended soley by the mediated pathway. Harm and Seidenberg demonstrate that the model's performance with homophones matches the findings from a number of empirical studies (Jared & Seidenberg, 1991; Lesch & Pollatsek, 1993; Van Orden, 1987; see also Van Orden & Kloos, this volume).

Conclusion

Connectionist models instantiate a set of computational principles that are intended to approximate the core properties of neural computation. Early efforts to apply these models to reading employed localist representations for words and hand-specified connection weights. More recent efforts have focused on learning internal distributed representations that effectively mediate the interaction of orthographic, phonological and semantic information. Because such systems lack word-specific representations and separate pathways for regular versus irregular items, they stand in sharp contrast to traditional dual-route theories of word reading. Moreover, exisiting models are still limited in the size and diversity of vocabulary they handle and the range of empiricial issues they address. Nonetheless, these system illustrate how a common computational framework can provide insight into reading acquisition, normal skilled reading, patterns of reading impairment following brain damage, and even possible approaches to remediation of developmental (Harm, McCandliss, & Seidenberg, 2003) and acquired (Plaut, 1996) deficits.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science*, *9*(2), 147-169.

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory and Cognition*, *10*, 565-575.

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review*, *105*(4), 678-723.

Behrmann, M., Plaut, D. C., & Nelson, J. (1998). A literature review and new data supporting an interactive account of letter-by-letter reading. *Cognitive Neuropsychology*, *15*, 7-51.

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*(3), 432-446.

Bullinaria, J. A. (1997). Modeling reading, spelling, and past tense learning with artificial neural networks. *Brain and Language*, *59*(2), 236-266.

Carr, T. H., Davidson, B. J., & Hawkins, H. L. (1978). Perceptual flexibility in word recognition: Strategies affect orthographic computation but not lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 674-690.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (p. 151-216). New York: 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*(4), 589-608.

Coltheart, M., Patterson, K., & Marshall, J. C. (Eds.). (1980). *Deep dyslexia.* London: Routledge & Kegan Paul.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature, 337*, 129-132.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton-Mifflin.

Frawley, W. (1992). *Linguistic semantics.* Hillsdale, NJ: Lawrence Erlbaum.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518-565.

Harm, M. W., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Study of Reading*, *7*(2), 155-182.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491-528.

Harm, M. W., & Seidenberg, M. S. (in press). *Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes.*

Hebb, D. O. (1949). *The organization of behavior.* New York: John Wiley & Sons.

Hillis, A. E., & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain, 114*, 2081-2094.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann Machines. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (p. 282-317). Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74-95.

Jared, D., & Seidenberg, M. S. (1991). Does word identification proceed from spelling to sound to meaning? *Journal of Experimental Psychology: General*, *120*(4), 358-394.

Kawamoto, A. (1988). Distributed representations of ambiguous words and their resolution in a connectionist network. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence.* San Mateo, CA: Morgan Kaufmann.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing approach. *Journal of Memory and Language, 32*, 474-516.

Kawamoto, A. H., Kello, C., & Jones, R. (1994). Locus of the exception effect in naming. In *Proceedings of the 35th Annual Meeting of the Psychonomic Society* (p. 51). St. Louis, MO.

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

Lesch, M. F., & Pollatsek, A. (1993). Automatic access of semantic information by phonological codes in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 285-294.

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

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, *27*, 213-234.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, *23*, 1-44.

McClelland, J. L., & Johnston, J. C. (1977). The role of familiar units in perception of words and nonwords. *Perception and Psychophysics*, *22*, 243-261.

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*(5), 375-407.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235-312.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*(1), 113-148.

Mozer, M. C. (1983). Letter migration in word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 531-546.

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach.* Cambridge, MA: MIT Press.

Mozer, M. C., & Behrmann, M. (1990). On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience*, *2*(2), 96-123.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895-938.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, *23*(4), 443-467.

Patterson, K., Coltheart, M., & Marshall, J. C. (Eds.). (1985). *Surface dyslexia.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Peterson, C., & Anderson, J. R. (1987). A mean field theory learning algorithm for neural nets. *Complex Systems*, *1*, 995-1019.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 291-321.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language*, *52*, 25-82.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes*, *12*, 767-808.

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science*, *23*(4), 543-568.

Plaut, D. C., & McClelland, J. L. (2000). Stipulating versus discovering representations [commentary on M. Page, Connectionist modelling in psychology: A localist manifesto]. *Behavioral and Brain Sciences*, *23*(4), 489-491.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (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*(5), 377-500.

Reicher, G. M. (1969). Perceptual recognition as a functionof meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 274-280.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533-536.

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

Saffran, E. M., Bogyo, L. C., Schwartz, M. F., & Marin, O. S. M. (1980). Does deep dyslexia reflect right-hemisphere reading? In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia* (p. 381-406). London: Routledge & Kegan Paul.

Seidenberg, M. S. (1992). Beyond orthographic depth: Equitable division of labor. In R. Frost & K. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (p. 85-118). Amsterdam: Elsevier.

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

Seidenberg, M. S., Petersen, A., MacDonald, M. C., & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 48-62.

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L., & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1177-1196.

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 Behaviour*, *23*, 383-404.

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

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, *3*, 9-44.

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

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory and Cognition*, *15*, 181-198.

Van Orden, G. C., & Goldinger, S. D. (1994). Interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1269-1291.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, *97*(4), 488-522.

Warrington, E. K. (1981). Concrete word dyslexia. *British Journal of Psychology*, *72*, 175-196.

Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time course and decision criteria. *Memory and Cognition*, *13*, 557-572.

Wheeler, D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*, 59-85.

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*(4), 490-501.

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