In Eden, G. F. (Ed.) (2015). *The Wiley handbook on the cognitive neuroscience of dyslexia.* Hoboken, NJ: John Wiley & Sons, Ltd.

Neurocomputational Principles of Reading

David C. Plaut
Department of Psychology
Center for the Neural Basis of Cognition
Carnegie Mellon University

Understanding how children learn to read, why some children fail at the task, and what might be done to help them succeed, is a challenge that is both extremely important and extremely difficult. The importance stems from the fact that text is a fundamental form of cultural and interpersonal communication and the primary vehicle by which most other forms of education take place. The difficulty of the challenge stems from the fact that reading acquisition depends on the interplay between two things that are themselves complex: the nature of the problems to be solved, and the nature of information processing and learning in the brain.

The Problem of Reading Acquisition

As a task, reading is particularly challenging as it involves the real-time derivation and coordination of many different types of information: orthographic, phonological, semantic and linguistic. Moreover, because writing is a relatively recent cultural invention—perhaps only about 5,400 years old—the brain cannot have evolved reading-specific mechanisms. Rather, reading must be learned on top of more general mechanisms for visual recognition and for spoken language (Dehaene & Cohen, 2007). Indeed, writing systems take advantage of, and are constrained by, the general biases of the visual system: the distribution of line junctions used in the world's orthographies corresponds quite closely to the distribution that occurs in natural scenes (and to which the visual system is tuned; Changizi, Zhang, Ye, & Shimojo, 2006), and many of the properties of eye-movements in reading derive from principles that govern them in visual object processing more generally (Reichle, Pollatsek, & Rayner, 2012).

Writing systems also vary widely in the granularity of the "units" they employ and how systematically these units map onto aspects of the phonology, semantics, and grammatical structure of the spoken language (Frost, 2012). Alphabetic scripts like English stick closely to phonology, using small units like single letters and multi-letter graphemes (e.g., CH, TH) to correspond to individual phonemes (albeit typically with some degree of inconsistency; cf. YACHT). Other scripts use larger and more complex orthographic elements (e.g., radicals in Chinese) to convey larger units in phonology (e.g., syllables) as well as broad semantic information. These differences in writing systems are fundamental to understanding reading acquisition: although a given child may be faced with learning only a single script, the cognitive and neural mechanisms that he or she brings to bear must be sufficiently flexible to learn *any* script (Seidenberg, 2011).

Reading acquisition is further complicated by being a slow, incremental process. Over a number of years, each step of learning to read is built on—and thus, for better or worse, influenced by—the solutions to previous, simpler problems. For example, the quality of preliterate phonological representations strongly impacts the efficacy of early learning of grapheme-phoneme correspondences in alphabetic scripts (Bradley & Bryant, 1983). As a result, weakness in basic visual or phonological processes can snowball into pervasive reading difficulties (Torgesen, Wagner, & Rashotte, 1994). Moreover, unlike most visual and linguistic skills, reading is typically acquired through explicit instruction, with the specific nature and efficacy of this instruction varying widely across cultures and by school district, and only very rarely is such instruction tailored to individual strengths and weaknesses (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

Reading is fundamentally about extracting meaning from text. However, given that children come to the task of learning to read with considerable knowledge of spoken language, they accomplishing this goal through a mixture of two broad strategies. On the one hand, they can learn to map visual representations of text onto their meanings directly, just as they have learned to recognize and understand other visual objects in their environment. On the other hand, insofar as the script they are learning contains units that reliable convey phonological information, readers can first learn to "decode" this information to derive the pronunciations of written words and then use their pre-existing spoken language knowledge to map these pronunciations onto meaning. In-

To appear in Eden, G. (Ed.), *Wiley Handbook on the Cognitive Neuroscience of Developmental Dyslexia.* Hoboken, NJ: John Wiley & Sons, Ltd.

Correspondence concerning this chapter may be sent to Dr. David Plaut, Department of Psychology, Carnegie Mellon University, Pittsburgh PA 15213–3890; email: plaut@cmu.edu.

deed, in most models of word reading, these strategies involve structurally separate routes or pathways: a *direct* pathway from orthography to semantics (typically via some type of lexical representation), and an *indirect* pathway from orthography first to phonology and then to semantics.[1] Note that the relative ease of acquisition of these pathways varies not only with individual differences in visual, phonological and semantic representations but also with the nature of the script being learned. Alphabetic scripts that convey detailed phonology but very little semantics will strongly favor the indirect pathway, whereas non-alphabetic scripts with larger units that convey both general phonological and semantic information will encourage greater involvement of the direct pathway.

In summary, successful reading acquisition depends on a highly complex interaction between the quality of the reader's preliterate representations, the nature of the script or scripts being learned, and the nature and extent of reading instruction and practice.

Computational Modeling

Given these complexities, any account of how and why most but not all children come to be skilled readers is itself going to be necessarily complex. In light of this, many researchers have turned to computational modeling as a way of supporting theory development. In this context, computational modeling is primarily a means of specifying the nature of reading-related information processing, representation, and learning with greater precision than is possible with verbal description alone. This increased precision assists both in verifying that the theory actually gives rise to the patterns of behavior it is intended to account for, as well as in generating more specific, testable empirical predictions to be examined in future work.

Computational modeling is, however, not without its pitfalls. In particular, any given implementation necessarily includes details that are not intended as theoretical claims but are introduced merely as simplifications, approximations, or computational conveniences. As a result, in comparing the behavior of a model with empirical findings, there is always an issue of the degree to which its successes (or failures) are due to properties of the underlying theory itself or due to incidental aspects of the implementation (including parameters that allow a model to overfit data; Seidenberg & Plaut, 2006).

One approach to addressing this issue is to evaluate multiple variants of a given model, varying theoretically irrelevant aspects or parameters, to assess the robustness of the model's theoretically relevant behavior (Pitt, Kim, Navarro, & Myung, 2006). Another, complementary approach is to further constrain the model by introducing additional considerations that go beyond the current domain of interest. This can involve developing the model using the same computational principles that have been applied successfully to many other problems and domains (thereby contributing to a much broader theoretical account of cognitive behavior). It can also involve incorporating constraints from other levels of analysis of the system, including its underlying neural basis. These two types of constraint are combined in an approach to computational modeling known variously as connectionist modeling, parallel distributed processing, or artificial neural network modeling (Elman et al., 1996; McLeod, Plunkett, & Rolls, 1998; McClelland, Rumelhart, & the PDP Research Group, 1986; O'Reilly & Munakata, 2000; Rogers & McClelland, 2014; Rumelhart, McClelland, & the PDP Research Group, 1986), which has seen a recent resurgence under the rubric of *deep learning* (Hinton, 2007; Hinton & Salakhutdinov, 2006; Schmidhuber, 2015).

Principles of Neural Computation

Connectionist/neural network modeling is an attempt to capture key principles of neural computation in a way that is simplified but still effective and informative for understanding computation in the brain. The core idea behind this approach is that cognitive processing takes the form of cooperative and competitive interactions among a large number of simple, neuron-like processing units. Understanding the relevance of such a system for modeling cognitive processing in general, and reading acquisition in particular, involves considering issues related to processing, network architecture, representation, and learning.

Processing

The units in neural network are intended to approximate individual neurons, but not in their full complexity. Rather, units aim to capture, in a simple a manner as possible, the basic information-processing characteristics of neurons while abstracting away the specific details of their biology. The typical assumption is that the state of a given neuron can be approximated by something like its instantaneous firing rate or *activation* level, and that the influence of one neuron on another is limited to a positive (excitatory) or negative (inhibitory) factor or *weight* applied to this activation. (A single connection weight summarizes the combined influence of all of the synapses that one neuron would make on another.) At any instant, the total net input to a given unit is the sum of positive- and negative-weighted activations of units from which it receives connections. In this way, input across positive weights increases a unit's net input (excitation), whereas input across negative weights decreases it (inhibition), with

[1] In some models (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2007, 2010), the indirect pathway itself has separate lexical and sublexical components, whereas in others (e.g., Harm & Seidenberg, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989) it consists of a single, uniform mechanism.

the degree of excitation or inhibition determined both by the magnitude of the weight and the activation level of the corresponding sending unit. The receiving unit's own activation is then set as a smooth, monotonic function of this total input (typically with a saturating nonlinearity).

Although drastically simplified relative to the operation of real neurons, this formulation of individual units in a neural network has fundamental implications for how the network responds to familiar inputs and generalizes to novel ones. To understand this, consider a single unit receiving weighted connections from a set of other units. The pattern of positive and negative weights on the connections determines how the receiving unit responds to any given pattern of activity over the sending units. (Learning involves adjusting these weights based on performance feedback, as discussed below.) The receiving unit will respond maximally if there are strong activations along positive-weighted connections and little if any activations across negatively-weighted connections. Conversely, the unit will be strongly inhibited if the reverse is true. Thus, the weights constitute something like a template that defines the optimal input pattern for the unit, as well as the degree to which it should respond to any given suboptimal input. In particular, if the input is similar but not identical to the optimal one, the unit's response will be reduced somewhat but might still be quite strong. This is because the weighted sum that determines its net input (and hence its activation level) will have most of the same terms as in the optimal case (due to the similarity of the two inputs). Note that it doesn't matter whether a given input is novel or familiar; each unit in the network responds as a function of the degree of match between its current input and its optimal one (as determined by learning). In this way, if the network has learned to respond appropriately to a given input, it will tend to respond similarly (generalize) to similar inputs.

The cooperation and competition among units (via positive and negative weights, respectively) cause the network to settle to a stable pattern of activity in response to a given input (typically provided by fixing the states of some units or providing them with an external contribution to their net inputs). This pattern can be understood as the network's interpretation of the input, including its response (over a designated set of output units). Such stable patterns are sometimes referred to as *attractors* because interactions among units cause similar patterns to move towards (be attracted by) the nearest attractor pattern, thereby cleaning up noise or other irrelevant variation among the unit activations.

Network Architecture

Only the simplest neural networks consist of a single, undifferentiated group of interconnected units. More typically, units are organized into layers, with earlier layers providing bottom-up input to later layers and (in some models) later layers providing top-down feedback to earlier layers. This fits with the broad hierarchical organization of neocortex (Felleman & Van Essen, 1991). The input to the entire system is provided as a pattern of activity over the first layer, and the output or response is generated over the last layer. Between the input and output layers are one more more intermediate or *hidden* layers that play a critical role in learning complex input-output mappings (as described in the next subsection).

In some models, the layers are not ordered in a hierarchy, but rather each represents a different type of information (along with hidden units that mediate their pairwise interactions). An example is the so-called "triangle model" of word reading (Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989) in which separate groups representing orthography, phonology, and semantics interact with each other via three groups of hidden units (thus forming a triangle). When, say, orthographic input for a word is provided to the network (by fixing the states of the orthographic units), unit interactions throughout the network cause it to settle into a global stable pattern that includes patterns over the semantic and phonological layers. In this way, the network as a whole generates or computes the meaning and pronunciation of the given word from its orthography.

Representation

It's generally not possible to model the entirety of perceptual, cognitive, and motor processing from retinotopic input to motor output. Rather, a model captures only a particular, theoretically interesting subset of the system, with earlier processes approximated by the nature of the input to the model, and later processes approximated by how output activations of the model are interpreted and related to behavior. This raises the question of exactly how entities such as words and objects—and their corresponding behaviors—should be represented in terms of unit activations.

The simplest possibility would be to assign a single, dedicated unit to each familiar entity (Bowers, 2009; Page, 2000), but such a *localist* representation has some significant drawbacks (Plaut & McClelland, 2000, 2010). Perhaps chief among these—particularly for input representations—is that localist units fail to capture the similarities among entities that, as just discussed, are the basis for generalization. The natural alternative, known as *distributed* representation, is to encode different entities as alternative patterns of activity over the same set of units, such that the degree and nature of the overlap among representations captures the degree and nature of the similarities of the corresponding entities. Distributed representations can, of course, vary in many ways, including their overall sparseness (i.e., the proportion of units typically activated by any given entity) and perplexity (i.e., the degree of dissimilarity among the entities to which a given unit contributes), with some extreme variants (very sparse, low-perplexity) behaving somewhat similarity to localist representations (O'Reilly & McClelland, 1994).

To be clear, the goal in designing distributed input or output representations for a model is to capture the relevant similarities among entities that result from unimplemented portions of the system, and as such benefit from the incorporation of constraints from empirical findings that bear on the characteristics of those parts of the system. Occasionally, in attempting to achieve this goal, researchers use units which are, themselves, localist representations of finer-grained entities—such as encoding whole word input in terms of units corresponding to individual letters (McClelland & Rumelhart, 1981; Plaut et al., 1996). This is often a computational convenience for achieving the appropriate similarities among higher-level entities and need not entail a theoretical commitment to localist representations of the lower-level entities (in a context in which they were the focus of interest). In general, it is important to keep in mind that a representation is localist or distributed only relative to a particular set of entities.

Learning

If input units were connected directly to output units, knowledge of how inputs map to outputs—or, more precisely, how input features map to output features—could be encoded in terms of the positive- or negative-valued weights on connections between them. In such a two-layer network, adjusting the weight values to achieve a given mapping is relatively straightforward: if, for a given input, a particular output unit is not sufficiently active, increase positive weights and decrease negative weights from active input units (or the reverse, if the unit it too active). Unfortunately, such networks can learn only fairly simple, similarity-preserving mappings (Minsky & Papert, 1969). In order to learn more complex mappings, like those required to comprehend and pronounce written words, one or more intermediate or hidden layers are required to mediate between inputs and outputs. Although such units do not themselves have specified target output activations, their incoming weights (and, hence, their activations) can be adjusted based on the impact such adjustments have on performance (i.e., by gradient descent in an error measure; Rumelhart, Hinton, & Williams, 1986). In this way, the network can gradually learn distributed internal representations over hidden units that are effective in mapping inputs to outputs—or in coordinating the interactions among multiple types of information—based on performance feedback. The similarities among learned hidden representations tend to "split the difference" between the similarities defined by the input representations and the similarities defined by the output representations.

Generalization is aided by preserving the general tendency to map similar inputs to similar outputs, as in a linear mapping. A completely linear multilayer network would be unhelpful, however, as it is no more powerful than a two-layer network (Minsky & Papert, 1969). Using a unit function with a mostly linear response for small and moderate net input (positive or negative), but a saturating nonlinearity for larger input, provides a good compromise between the generalization abilities of a linear system and the computational power of a nonlinear system. In particular, when weights in the network are small (e.g., early on in learning), the network will behave mostly linearly. If such a linear system is sufficient to solve the task, the network will perform well and also generalize optimally. Insofar as some aspects of the task require nonlinearities—that is, exceptional aspects in which input similarity does not correspond to output similarity (e.g., YACHT pronounced "yot" instead of "yatched")—the network will have to grow weights large enough to drive some units into the nonlinear range of their input-output function, but this will take time and only happen if necessary to achieve good performance. In this way, the network will eventually learn to cope with idiosyncratic aspects of a task while remaining as close to linear as possible (to aid generalization).

There are, of course, many contexts in which it is necessary to learn idiosyncratic information rapidly, such as when learning the name of a new acquaintance or remembering where you parked your car in a parking lot. The slow integration of idiosyncratic and systematic knowledge, as just described, is ill-suited to such demands. In fact, there is compelling evidence that the brain employs a separate, hippocampal-based system for rapid learning of arbitrary combinations of information (i.e., so-called *episodic* knowledge; Tulving, 1983) in a way the complements the slower, similarity-based learning in neocortex (McClelland, McNaughton, & O'Reilly, 1995). On this account, rapid hippocampal learning not only supports immediate performance directly but also provides additional off-line training of neocortex to help consolidate new information with long-term knowledge, thereby giving rise to cortical representations that most effectively capture the underlying structure of the domain.

Summary

Connectionist/neural-network modeling provides a computational framework that attempts to capture the core principles of processing, representation and learning in the brain. It necessarily abstracts from many specific details of the structure and function of individual neurons, and how large ensembles of them interact to support complex behavior. Nonetheless, the general success of the framework in accounting for central aspects of acquisition and skilled performance in a variety of perceptual, cognitive, and motor domains (McClelland et al., 1986; McLeod et al., 1998; O'Reilly & Munakata, 2000; Rumelhart, McClelland, & the PDP Research Group, 1986) as well as patterns of impaired performance in acquisition and following brain damage (Elman et al., 1996; Plaut & Shallice, 1993) suggests that the principles it embodies provide important insights into learning and processing in the brain.

In the context of typical and atypical reading acquisition, the most central issues relate to 1) the use of distributed representations in which pattern overlap over different groups of units captures different types of similarities (orthographic, phonological, semantic); 2) the relative ease of learning a systematic mapping (in which similar inputs map to similar outputs; e.g., orthography-to-phonology in English) compared to an unsystematic mapping (in which input similarity is unrelated to output similarity; e.g., orthography-to-semantics); and 3) differences among languages and scripts in the relative systematicity of mapping orthography directly to semantics versus mapping orthography via phonology to semantics.

Neurocomputational Models of Reading

The remainder of this chapter provides a survey of models of various aspects of normal and impaired reading that are based on the principles of neural computation just described. The survey is necessarily selective but aims to illustrate the breadth and usefulness of the overall approach. By necessity, any given model instantiates only a part of the entire processing system and only a subset of the relevant behavioral data. Nonetheless, collectively, the set of models supports a coherent theory of the neural basis of reading—including its acquisition, skilled performance, and breakdown following brain damage—in part because they each are based on largely the same neurocomputational principles.

Orthographic processing

One of the earliest and most influential models of orthographic processing is the Interactive Activation (IA) model of letter and word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). The model consists of three layers of units: 1) letter-feature units at each of four possible positions, corresponding to the various strokes that make up letters and constituting the bottom-up input to the model; 2) letter units at each of the four position, with one unit for each possible letter; and 3) word units at the top, with one unit for each known word (i.e., a localist representations of words). Each letter unit at a position (e.g., T in position 1) is supported by the features it contains (via bottom-up excitatory connections), competes with the other letters at that position (e.g., F in position 1, via bidirectional inhibitory weights), and cooperates with words containing the letter at that position (e.g., TIME, via bidirectional excitatory weights). At the word level, alternative words compete with each other (via inhibitory weights) and support the letters they contain (via excitatory weights).

Although the organization of the layers is hierarchical (broadly in keeping with visual cortical areas; Felleman & Van Essen, 1991), processing does not proceed in stages from features to letters to words. Rather, processing throughout the network is graded and interactive, so that partial activation at each level contributes to and constrains the partial activation at the other levels to which it is connected, in both a feedforward and feedback manner. As a result, activations at both the letter level and the word level mutually constrain each other to settle to patterns of activity that are maximally consistent with each other and with the featural input. When featural information is weak, noisy or missing, the resulting partial letter activation can be cleaned-up by top-down support from partially consistent word units. In this way, the model provides a detailed explanation for a variety of effects related to the word superiority effect (Reicher, 1969; Wheeler, 1970), in which the perception of letters is better when embedded in words than when embeded in consonant strings or non-letter symbols.

The fact that the IA model remains highly successful in explaining phenomena within its restricted domain (as does its spoken-word analog, the TRACE model; Allopenna, Magnuson, & Tanenhaus, 1998; Magnuson, Mirman, & Harris, 2012; Mayor & Plunkett, 2014; McClelland & Elman, 1986) suggests that it captures something fundamental about perceptual processing. Nonetheless, certain specific aspects of the IA model's design are clearly limited or wrong in detail, including the restriction to four-letter words, the use of strict position-specific coding of letters, the use of localist word representations, and the lack of learning in the model. Many of these limitations have been addressed in subsequent modeling efforts.

With regard to orthographic encoding per se, a number of empirical results, including strong priming of words from transposed-letter primes (JUGDE-JUDGE; Kinoshita & Norris, 2009; Perea & Lupker, 2004; Schoonbaert & Grainger, 2004), indicate that word length and letter-position information must be represented more flexibly than position-specific slots. A number of specific proposals have been made in this regard, including the SOLAR model (C. J. Davis, 1999; C. J. Davis & Bowers, 2006), the SERIOL model (Whitney, 2001, 2008, 2011) and the Overlap model (Gomez, Ratcliff, & Perea, 2008). By and large, all of these models succeed at accounting for positional flexibility but—with the possible exception of the Overlap model—they do so at the expense of building in constraints that hold within Indo-European scripts but not universally. Indeed, Lerner, Armstrong, and Frost (2014) have shown recently that positional flexibility, as reflected by transposed-letter effects, is not an intrinsic property of orthographic representations but varies cross-linguistically as a function of the statistical structure of different scripts. Consequently, they argue—and support with a series of connectionist simulations—that a comprehensive account of orthographic representations cannot simply stipulate how such representations are organized but must be based on a theory of how they are learned in response to reading experience.

The remaining major limitations of the IA model—namely, the use of localist word representations and the lack of learning—have been addressed largely in the context of models that learn to map orthography to phonology and/or semantics in understanding and pronouncing written words.

Mapping orthography to phonology

In spite of the fact that the primary purpose of reading is comprehension, an inordinate amount of theoretical attention has been paid to processes involved in reading aloud. This is partly because overt pronunciations are easier to measure, but also because the relationship between spelling and sound in English (and, to varying degrees, in other languages/scripts) is *quasiregular* in that it is largely systematic but admits many exceptions which exhibit subregularities among them (Seidenberg & McClelland, 1989). Quasiregular structure is common across a number of linguistic and non-linguistic domains, including the English past-tense (Rumelhart & McClelland, 1986; Seidenberg & Plaut, 2014) and the organization conceptual knowledge (McClelland et al., 1995; Rogers & McClelland, 2004), and it evokes a recurring theoretical debate on the relationship between systematic and idiosyncratic knowledge and how best to capture each.

On the one hand, strong intuitions suggest that systematic knowledge is best expressed in terms of abstract, explicit rules that are insensitive to any variation outside the scope of the rule and hence support effective generalization over such variation (Pinker, 1999). For instance, systematic English spelling-sound correspondences can be expressed by grapheme-phoneme correspondence (GPC) rules (e.g., E ⇒ "eh"; Venezky, 1970). A set of such rules can be used to piece together or "sound out" most English words (e.g., SET) and also generalize to pronounceable nonwords (e.g., SEK). However, for about 20% of English monosyllabic words (e.g., SEW), such rules give the wrong result ("sue" instead of "so") and so separate, word-specific knowledge is needed to override the rules to pronounce such *irregular* or *exception* words correctly. These considerations lead to so-called *dual-route* models of word reading (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart et al., 2001; Perry et al., 2007, 2010) in which systematic and idiosyncratic knowledge are represented and processed separately (and integrated only at the level of phonological responses). Although these models differ in the specific nature and implementation of GPCs within the sublexical pathway, they all use a variant of the IA model as the lexical pathway that enforces word-specific knowledge, and hence inherit the limitations of that model.

Distributed connectionist modeling provides an alternative to explicit rules and the strict separation of systematic and idiosyncratic knowledge (McClelland & Rumelhart, 1985; Rumelhart & McClelland, 1986). In learning to map patterns of inputs to patterns of outputs, connectionist networks develop internal representations that capture not only how parts of the output depend on parts of the input, but also the extent to which parts of the output are *independent* of other parts of the input. Thus, in mapping spelling to sound, the network will learn that an initial B reliably predicts initial phoneme /b/ in the output, but also that initial B ⇒ /b/ doesn't depend on anything else in the input. As a result, other inputs that contain B (e.g., BEK) will correctly activate /b/ regardless of whether or not they are familiar, thereby supporting effective generalization. At the same time, the network can be sensitive to the entire input if required by some aspect of the output (e.g., the vowel in SEW). In this way, the same system can learn idiosyncratic knowledge for exception items while still supporting effective generalization to novel inputs.

Following Rumelhart and McClelland's (1986) work on the English past tense, Seidenberg and McClelland (1989) trained a network to map from orthography to phonology for about 2800 English monosyllabic words.[2] The network succeeded in accounting for a wide range of empirical results concerning the interaction of word frequency and spelling-sound consistency using real-valued error (for correct pronunciations) as a proxy for response latency. However, the model was not as accurate as skilled readers at naming pronounceable nonwords, particularly those with somewhat unusual spellings. In follow-up work, Plaut et al. (1996) showed that the limitations generalization of the Seidenberg and McClelland (1989) model stemed from the use of poorly structured input and output representations (derived from Rumelhart & McClelland, 1986); when representations based on graphemes and phonemes were used, networks can learn to pronounce both regular and exception words correctly while still generalizing to nonwords as well as skilled readers. Moreover, they can account for effects of frequency and consistency on naming latencies in terms of actual processing time in generating stable phonological output.

Mapping orthography to semantics

In contrast to the highly (if only partially) systematic mapping between orthography and phonology in English, the mapping from the surface forms of words to their meanings is largely unsystematic (apart from morphological regularities like TEACH-TEACHER). That is, words that are orthographically or phonologically similar (e.g., PEACH-BEACH) are no more likely to be semantically related than words whose forms are dissimilar (e.g., PEACH-APPLE).

[2] The network was also trained to regenerate the orthographic input over a separate set of output units, for use in modeling lexical decision performance. However, its ability to accomplish this task (as well as follow-up work by Plaut, 1997) is less directly relevant to the current issues at hand.

Given that networks are intrinsically sensitive to similarity, this makes word comprehension much more difficult to learn than word pronunciation. This difficulty has two important implications. The first is that the system will rely on rapid, hippocampal-based learning to support effective performance during the extended time it takes for knowledge of a new word meaning to be integrated effectively with the rest of a person's vocabulary (M. H. Davis & Gaskell, 2014). The second is that, insofar as reliable phonology-semantics mappings have been established prior to reading acquisition, the system will rely on the easier-to-learn orthography-to-phonology mapping to derive a pronunciation that can then be mapped to semantics using this preexisting language knowledge (Frost, 1998). Moreover, once this is accomplished, the derived meaning can form the basis for training the direct orthography-to-semantics mapping (Share, 1995).

In general, successful reading will involve the coordinated involvement of both the direct and indirect (phonologically mediated) pathways. In fact, for items like homophones (e.g., STAKE/STEAK), a contribution of the direct pathway is needed to drive the appropriate meaning. Harm and Seidenberg (2004) carried out an extensive series of connectionist simulations exploring the learned division-of-labor between direct and indirect contributions to mapping orthography to semantics, and showed how such a system can account for various findings concerning the processing of homophones and pseudohomophones (i.e., a nonword with a lexical pronunciation, such as BRANE).

More recent work has examined the processing dynamics of the direct orthography-semantics pathway in more detail. Armstrong and Plaut (2008, 2014) attempted to account for effects of semantic ambiguity—in particular, how words with multiple related meanings are processed more quickly than unambiguous controls in tasks like lexical decision, whereas words with unrelated meanings are processed more slowly than controls (see also Rodd, Gaskell, & Marslen-Wilson, 2002). Their account is based on the idea the early dynamics within the network is dominated by cooperative interactions among features shared by related meanings, whereas later dynamics are dominated by competitive interactions among the non-overlapping features of unrelated meanings. Their model accounts for both early and late effects by incorporating additional neurophysiologically motivated constraints (i.e., using separate populations of excitatory and inhibitory units, and restricting between-layer communication to excitation only).

Laszlo and Plaut (2012) used a network with essentially the same additional constraints to account for properties of the N400 evoked response potential (ERP) component, commonly interpreted as reflecting semantic integration (Kutas & Federmeier, 2011). In the model, the N400 reflects transient semantic over-activation (during Armstrong & Plaut's early cooperative phase). Laszlo and Plaut showed that this semantic measure accounts for the seemingly paradoxical findings of a strong effect of orthographic neighborhood size (words and pseudowords vs. acronyms and consonant strings) but little effect of meaningfulness (words and acronyms vs. pseudowords and consonant strings) on single-item N400 magnitudes (Laszlo & Federmeier, 2011). Laszlo and Armstrong (2014) extended to model to address repetition priming effects by introducing a decay function that approximates the temporal dynamics of cortical post-synaptic potentials.

Acquired dyslexia

Important constraints on the organization and operation of the reading system have come from detailed studies of reading impairments caused by various types of brain damage. Although some modeling efforts have been directed at accounting for peripheral acquired dyslexias, including neglect dyslexia (Mozer & Behrmann, 1990) and pure alexia (Plaut, 1999), most theoretical attention has been focused on understanding central acquired dyslexias—particularly surface dyslexia and deep/phonological dyslexia.

Surface dyslexia is marked by largely intact reading of regular words and nonwords but the production of "regularization" errors to exception words, particularly those of low frequency (e.g., SEW ⇒ "sue"). The latter items are difficult for networks to master, both because they are trained less often and also because their mapping conflicts with those of orthographically similar words (FEW, GREW, KNEW, etc.). It thus makes sense, on a distributed network account, that low-frequency exception words are among the last to be acquired (Backman, Bruck, Hébert, & Seidenberg, 1984) and the most vulnerable to damage (Patterson, Marshall, & Coltheart, 1985). However, Patterson, Seidenberg, and McClelland (1989) had only limited success in modeling surface dyslexia by damaging the Seidenberg and McClelland (1989) model. Plaut et al. (1996) obtained similar results when damaging purely orthography-to-phonology models, but had much greater success when incorporating semantics into a full version of the triangle model. In particular, a "phonological" (orthography-to-phonology) pathway trained in conjunction with a gradually increasing approximation of a "semantic" (orthography-to-semantics-to-phonology) pathway accounted for skilled performance when intact. Underlying this skilled performance was a graded division-of-labor between the pathways in which the phonological pathway alone did not become fully competent at the items it finds most difficult—namely, low-frequency exception words. As a result, when the semantic contribution was gradually compromised (as an approximation to progressive semantic dementia; Graham, Hodges, & Patterson, 1994) the intact but progressively isolated phonological pathway provided a good match to varying levels of severity of surface dyslexia. Moreover, by also varying premorbid strength of the semantic pathway, Woollams, Lambon Ralph, Plaut, and Patterson

(2007) showed that the model could account for the distribution of effects exhibited by 100 observations over a large cohort of semantic dementia patients (see also Dilkina, McClelland, & Plaut, 2008).

Deep/phonological dyslexia, by contrast, involves relatively good word reading—including exception words—but much poorer nonword reading (often with a tendency to give visually related lexical errors; e.g., PINT ⇒ "print"). Deep dyslexia is generally thought to be the most severe form of phonological dyslexia (Crisp & Lambon Ralph, 2006; Friedman, 1996) and is marked by the additional occurrence of *semantic* errors (e.g., RIVER ⇒ "ocean"). Hinton and Shallice (1991) showed that semantic errors can arise from damage to an orthography-to-semantics network that develops "attractors" for word meanings: after damage, a visual input may be captured by the corrupted attractor for a nearby (semantically similar) words. Interestingly, like patients, the network also produces higher-than-chance rates of visual and mixed visual-and-semantic errors (e.g., TROUBLE ⇒ "terrible") because visually similar inputs (e.g., DOG, LOG) generate similar initial semantic activation, such that one of them can be incorrectly captured by the attractor for the other. Plaut and Shallice (1993) later extended the Hinton and Shallice account to address the full range of characteristics of deep dyslexia, including effects of imageability/concreteness.

Welbourne, Woollams, Crisp, and Lambon Ralph (2011) argued that a full account of both surface and phonological/deep dyslexia requires a consideration of spontaneous recovery following damage. First, they accounted for normal reading performance by developing a variant of the full triangle model with dense connectivity within each of orthography, phonology, and semantics but only sparse connectivity between these domains. When acute damage to phonology was followed by partial retraining, the model produced the core characteristics of phonological dyslexia (Crisp & Lambon Ralph, 2006), including the semantic errors of deep dyslexia if the initial damage was sufficiently severe. By contrast, when retraining was applied during progressive damage to semantics (analogous to semantic dementia), the model's behavior matched that of surface dyslexic patients (Woollams et al., 2007). The impressive breadth of coverage of this modeling effort argues strongly for the value of distributed connectionist modeling of both normal and impaired word reading, as well as the importance of considering postmorbid adaptation in interpreting the performance of brain-damaged patients.

Reading acquisition and developmental dyslexia

The emphasis that connectionist modeling places on learning makes it a natural framework within which to develop accounts of reading acquisition. Indeed, over the course of training, the Seidenberg and McClelland (1989) model showed sensitivity to graded degrees of spelling-sound consistency that mirrored empirical results for readers from second grade through high school (Backman et al., 1984). Moreover, halving the number of hidden units in the model (see also Plaut et al., 1996) produced exaggerated consistency effects and poor asymptotic performance on low-frequency exception words, as observed for children with the "surface" or "delayed" variant of developmental dyslexia (Manis, Seidenberg, Doi, McBride-Chang, & Peterson, 1996) However, the model was unable to account for "phonological" developmental dyslexia, marked by relatively good lexical reading but especially poor nonword reading.

Harm and Seidenberg (1999) explored whether the phonological subtype of developmental dyslexia might arise due to poorly structured preliterate phonological representations (Manis et al., 1997). When a network first learned to form phonological attractors for word forms (in the course of preliterate language acquisition), its subsequent reading acquisition, including both word and nonword reading, was fully successful. If, however, preliterate experience did not establish such attractors, nonword reading performance in particular was far poorer, mimicking phonological developmental dyslexia. By contrast, starting with normal phonological attractors but reducing the number of hidden units between orthography and phonology (as in Plaut et al., 1996; Seidenberg & McClelland, 1989) yielded the surface/delayed variant. More severe damage of either type, or a mixture of the two types of damage, yielded a mixed profile of impairment in which both exception words and nonwords suffered, as is observed in the majority of developmental dyslexic cases (Manis et al., 1996, 1999).

Powell, Plaut, and Funnell (2006) compared the initial trajectory of learning in the Plaut et al. (1996) orthography-to-phonology model against the correct performance and patterns of errors produced by children at two points during their first year of reading instruction. The original model produced poor performance on nonwords and fewer lexical errors compared to the children. However, when the model was trained in a way closer to actual reading instruction—including explicit grapheme-phoneme instruction, a gradually expanding training corpus, and vocabulary drawn from children's early reading material, the model provided a much closer fit to the empirical findings. The one remaining discrepancy—lower rates of lexical errors—were shown to be due to the lack of a full implementation of the triangle framework including semantics.

Modeling in other languages and scripts

The vast majority of neurocomputational modeling of reading has been applied to English. While this has produced important insights concerning how neural-like mechanisms learn and process quasiregular mappings, it has also led to a rather narrow view of the problems faced by children learning to read and, thus, to a failure to appreciate the generality of the system that solves these problems (Frost, 2012; Seidenberg, 2011). The degree of variability across the world's languages, in terms of their phonological, morphological, and syntactic organization as well as the scripts they employ to convey this information in print, is truly remarkable. Although a given child need not master all of this complexity, he or she must possess a cognitive and neural system that is capable to mastering any of it (and typically more than one). This degree of flexibility demands a general learning-based approach of the sort provided by distributed connectionist modeling (Lerner et al., 2014).

The visual complexity of orthographies varies widely across writing systems and strongly influences perceptual learning of graphemes, the initial stage of reading development (Chang, 2014). Chang, Plaut, and Perfetti (in press) carried out a series of computational simulations that examined the degree to which visual-orthographic complexity contributes to reading performance. They trained a hierarchical neural network, in which units had topographically constrained receptive fields of different sizes (as in visual cortex), to reconstruct the graphemes/characters from each of 130 different orthographies. Across writing systems, they found a strong, positive association between overall grapheme complexity of a script and network learning difficulty. In addition, they gathered data from same-difference judgments of pairs of graphemes drawn from six different orthographies (varying in complexity), made by native speakers (and readers) of eight different languages/writing systems (also varying in complexity), and posed exactly the same task to networks trained on each of those eight orthographies. Consistent with human performance, difficulty in processing graphemes by the networks was a function of complexity of the presented orthography itself as well as its relationship to the network's trained (native) orthography.

In considering how orthography maps to phonology and semantics, there have been a number of efforts to apply to triangle model to Indo-European languages other than English (e.g., Hutzler, Ziegler, Perry, Wimmer, & Zorzi, 2004). From the point of view of establishing the generality of the approach, however, perhaps the more interesting work has involved languages with scripts that are very different from English orthography, including Chinese (Yang, Shu, McCandliss, & Zevin, 2013) and Japanese (Ijuin, Fushimi, Patterson, & Tatsumi, 1999; Ueno, Ikeda, Ito, Kitagami, & Kawaguchi, 2014).

Skilled Chinese readers know upwards of 4,000 distinct characters, each of which is a complex visual pattern corresponding to a syllable or morpheme rather than a phoneme (as in alphabetic scripts). Most characters contain a component or "radical" that provides broad information about the character's meaning. Thus, the relative ease in mapping Chinese orthography to semantics versus phonology is very different than in English and, in fact, developmental reading deficits manifest very differently in these two writing system. Yang et al. (2013) simulated aspects of reading acquisition in Chinese and English using same the network for both writing systems—a variant of the triangle model. Due to the statistical differences between the scripts, and resulting differences in the division-of-labor between the phonological and semantic pathways, semantic or phonological deficits give rise to rather different patterns of impaired reading acquisition, each of which corresponds well to what is observed empirically. Moreover, the same results hold for a bilingual/biliteral network trained on both English and Chinese from the outset. Thus, different patterns of impaired reading acquisition across writing systems can be understood in terms of how a common reading architecture adapts to the different statistical structure among orthography, phonology, and semantics.

Japanese employs two types of script. Kanji consists of logographic characters borrowed from Chinese, whereas kana consists of syllabic characters and comes in two forms: hiragana for naturalized Japanese words, and katakana for foreign words/names. Thus, within the same language (and sometimes for the same words), kana provides a highly systematic mapping from orthography to phonology, whereas kanji is largely arbitrary in this regard, so that this mapping has no advantage over mapping orthography directly to semantics. Ijuin et al. (1999) developed a version of the triangle model that was trained both to comprehend and pronounce both kanji and kana words, and successfully simulated a number of effects seen in the reading performance of Japanese skilled readers. Damage to semantics produced a surface dyslexic pattern, in that performance on kana strings and on those kanji characters with consistent character-sound correspondences was much better than on kanji characters with atypical correspondences. By contrast, damage to phonology produced a phonological dyslexic pattern: reading of both kanji and kana words was much better than that of kana nonwords. Thus, the same principles that account for normal and impaired reading in English also account for analogous findings in Japanese. More recently, Ueno et al. (2014) have developed a larger-scale version of the triangle model as applied to Japanese in order to provide a better quantitative fit to skilled nonword reading.

Conclusions

Learning to read poses a difficult challenge for children, just as understanding how children learn to read poses a difficult challenge for researchers. Children are able to meet the challenge because their cognitive and neural systems embody particular computational principles for how orthographic, phonological and semantic information is learned, represented and processed. Accordingly, by developing simulations which instantiate these principles in working computational models of how these types of information interact, researchers can best understand how the process unfolds for a given child learning a given script of a given language.

This chapter discussed the principles that underlie distributed connectionist modeling, and reviewed the application of such models to reading acquisition and developmental dyslexia, normal skilled reading, and acquired dyslexia following brain damage. At the core of the models' success in modeling the relevant empirical phenomena is their intrinsic sensitivity to similarity among input patterns, their preference for learning systematic mappings that preserve similarity, and their ability to simultaneously learn idiosyncratic information when necessary (and only with enough practice). These properties explain why, within a given language, some aspects of reading are more difficult to learn than others (e.g., exception vs. regular words), why the models can successfully generalize their knowledge to novel inputs (i.e., pronounceable nonwords), and why certain patterns of impairment follow certain types of damage (e.g., surface dyslexia following semantic damage; phonological dyslexia following phonological damage). They also explain why, across languages, different writing systems give rise to different divisions-of-labor within the reading system—and, hence, different patterns of difficult in acquisition and following brain damage—due to the particular statistical structure among orthography, phonology and semantics.

Moreover, the same computational principles provide insight into many domains beyond reading per se, including other aspects of language processing (McClelland, St. John, & Taraban, 1989), visual object recognition (Krizhevsky, Sutskever, & Hinton, 2012), conceptual development and processing (Rogers & McClelland, 2004), learning and memory (McClelland et al., 1995), and executive functions (Botvinick & Cohen, 2014). In this way, the framework as a whole provides the promise for connecting reading research with findings from a broad range of other domains, thereby contributing to the development of a comprehensive theory of the neural basis of cognitive processing.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439.

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In *Proceedings of the 30th annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Armstrong, B. C., & Plaut, D. C. (2014, November). *Semantic ambiguity effects in lexical processing: A neural-network account based on semantic settling dynamics.* (Manuscript submitted for publication)

Backman, J., Bruck, M., Hébert, M., & Seidenberg, M. S. (1984). Acquisition and use of spelling-sound information in reading. *Journal of Experimental Child Psychology*, *38*, 114-133.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*, 1249-1285.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*, 220-251.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read: A causal connection. *Nature*, *301*, 419-421.

Chang, L. Y. (2014). *Visual orthographic variation and learning to read across writing systems* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh PA, USA.

Chang, L. Y., Plaut, D. C., & Perfetti, C. A. (in press). Visual-orthographic complexity in learning to read: Modeling learning across writing system variations. *Scientific Studies of Reading*.

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *The American Naturalist*, *167*(5), E117-E139.

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., 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.

Crisp, J., & Lambon Ralph, M. A. (2006). Unlocking the nature of the phonological-deep dyslexia continuum: The keys to reading aloud are in phonology and semantics. *Journal of Cognitive Neuroscience*, *18*, 348-362.

Davis, C. J. (1999). *The self-organizing lexical acquisition and recognition (SOLAR) model of visual word recognition* (Unpublished doctoral dissertation). University of New South Wales, Sydney, Australia.

Davis, C. J., & Bowers, J. S. (2006). Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 535-557.

Davis, M. H., & Gaskell, M. G. (2014). A complementary systems account of word learning: neural and behavioural evidence. *Proceedings of the Royal Society of London, Series B*, *364*, 3773-3800.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*, 384-398.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, *25*, 136-164.

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.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, *1*, 1-47.

Friedman, R. B. (1996). Recovery from deep alexia to phonological alexia. *Brain and Language*, *52*, 114-128.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, *123*(1), 71-99.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*, 263-239.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*, 577-601.

Graham, K. S., Hodges, J. R., & Patterson, K. (1994). The relationship between comprehension and oral reading in progressive fluent aphasia. *Neuropsychologia*, *32*(3), 299-316.

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. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662-720.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*, 528-434.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507.

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

Hutzler, F., Ziegler, J. C., Perry, C., Wimmer, H., & Zorzi, M. (2004). Do current connectionist learning models account for reading development in different languages? *Cognition*, *91*, 273-296.

Ijuin, M., Fushimi, T., Patterson, K., & Tatsumi, I. (1999). A connectionist approach to Japanese kanji word naming. *Psychologia*, *42*, 267-280.

Kinoshita, S., & Norris, D. (2009). Transposed-letter priming of prelexical orthographic representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1-18.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing 25.* Cambridge, MA: MIT Press.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621-647. doi: 10.1146/annurev.psych.093008.131123

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended event-related potential reading data. *Brain and Language*, *132*, 22-27.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*, 176-186.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of event-related potential word reading data. *Brain and Language*, *120*, 271-281.

Lerner, I., Armstrong, B. C., & Frost, R. (2014). What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language*, *77*, 40-58.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, M. McRae, & M. Joanisse (Eds.), *The cambridge handbook of psycholinguistics* (p. 76-103). Cambridge, UK: Cambridge University Press.

Manis, F. R., McBride-Chang, C., Seidenberg, M. S., Keating, P., Doi, L. M., Munson, B., & Petersen, A. (1997). Are speech perception deficits associated with developmental dyslexia? *Journal of Experimental Child Psychology*, *66*(2), 211-235.

Manis, F. R., Seidenberg, M. S., Doi, L. M., McBride-Chang, C., & Peterson, A. (1996). On the bases of two subtypes of developmental dyslexia. *Cognition*, *58*, 157-196.

Manis, F. R., Seidenberg, M. S., Stallings, L., Joanisse, M., Bailey, C., Freedman, L., ... Keating, P. (1999). Development of dyslexic subgroups: A one-year follow up. *Annals of Dyslexia*, *49*, 105-134.

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from TRACE simulations. *Journal of Memory and Language*, *71*, 89-123.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

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.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159-188.

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

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, *4*, 287-335.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. Oxford, UK: Oxford University Press.

Minsky, M., & Papert, S. (1969). *Perceptrons*. 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., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hipocampus*, *4*, 661-682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

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

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

Patterson, K., 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* (p. 131-181). London: Oxford University Press.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language*, *51*, 231-246.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested modeling and strong inference testing in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*, 301-333.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*, 106-151.

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

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57-83.

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., & McClelland, J. M. (2010). Locating object knowledge in the brain: Comment on BowersâĂŹs (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review*, *117*, 284-290.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377-500.

Powell, D., Plaut, D. C., & Funnell, E. (2006). Does the Plaut, McClelland, Seidenberg and Patterson (1996) connectionist model of single word reading learn to read in the same way as a child? *Journal of Research in Reading*, *29*, 229-250.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest Monograph*, *2*, 31-74.

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

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye-mind link. *Psychological Review*, *119*(1), 155-185.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245-266.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science*, *38*, 1024-1077.

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

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (p. 216-271). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85-117.

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes*, *19*, 333-367.

Seidenberg, M. S. (2011). Reading in different writing systems: One architecture, multiple solutions. In P. McCardle, J. Ren, & O. Tzeng (Eds.), *Across languages: Orthography and the gene-brain-behavior link* (p. 149-174). Baltimore, MD: Paul Brooke Publishing.

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

Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (p. 25-49). Hove, UK: Psychological Press.

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*, 1190-1228.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218.

Torgesen, J., Wagner, R., & Rashotte, C. (1994). Longitudinal studies of phonological processing and reading. *Journal of Reading Disabilities*, *27*, 276-286.

Tulving, E. (Ed.). (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

Ueno, T., Ikeda, K., Ito, Y., Kitagami, S., & Kawaguchi, J. (2014). Parallel vs. serial issues in reading aloud: Evidence for parallel processing from a computational model of Japanese kanji and kana nonword reading. In *Proceedings of the 36th annual conference of the cognitive science society* (p. 285-290). Mahweh, NJ: Lawrence Erlbaum Associates.

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

Welbourne, S. R., Woollams, A. M., Crisp, J., & Lambon Ralph, M. A. (2011). The role of plasticity-related functional reorganization in the explanation of central dyslexias. *Cognitive Neuropsychology*, *28*, 65-108.

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

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin and Review*, *8*, 221-243.

Whitney, C. (2008). Comparison of the SERIOL and SOLAR theories of letter-position encoding. *Brain and Language*, *107*, 170-178.

Whitney, C. (2011). Location, location, location: How it affects the neighborhood (effect). *Brain and Language*, *118*, 90-104.

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*, 316-339.

Yang, J., Shu, H., McCandliss, B. D., & Zevin, J. D. (2013). Orthographic influences on division of labor in learning to read Chinese and English: Insights from computational modeling. *Billingualism: Language and Cognition*, *16*, 354-366.