In G. Gaskell (Ed.), Oxford Handbook of Psycholinguistics. Oxford University Press, 2007, pp. 235-250.

CONNECTIONIST MODELS OF READING

Mark S. Seidenberg
Department of Psychology
University of Wisconsin-Madison

Running head: Connectionist models of reading

Department of Psychology
University of Wisconsin
Madison WI 53706 USA
seidenberg@wisc.edu

Much of what we know about the acquisition and use of language has resulted from close analyses of normal and disordered behavior. Since the late 1980s, another tool has been available: the building of connectionist computational models. These models have been extensively used in the study of reading: how children learn to read, skilled reading, and reading impairments (dyslexia). The models are computer programs that simulate detailed aspects of behavior. So, for example, a reading model might be taught to learn to recognize letter strings and compute their meanings or pronunciations. Such models provide a way of developing and testing ideas about how people read, in the service of developing a general theory. The purpose of this chapter is to provide an overview of connectionist models of reading, with an emphasis on the "triangle" framework developed by Seidenberg and McClelland (1989), Plaut et al. (1996), and Harm and Seidenberg (1999, 2004).

Basic Elements of Connectionist Models of Reading

The term "connectionism" refers to a broad, varied set of ideas, loosely connected (so to speak) by an emphasis on notion that complexity, at different grain sizes or scales ranging from neurons to overt behavior, emerges from the aggregate behavior of large networks of simple processing units. Our focus is on the parallel distributed processing (PDP) variety developed by Rumelhart, McClelland, and Hinton (1986). These models consist of large networks of simple neuron-like processing elements that learn to perform tasks such as reading words or recognizing objects. Our reading models were used to explore a more general theory of how lexical knowledge is acquired and used in performing several communicative tasks (speaking, listening, reading, writing), based on PDP principles. The reading models differ in detail but all conform to a common theoretical framework, the main elements of which will be summarized briefly.

Task Orientation

The models perform tasks such as computing the meanings or pronunciations of words. The goal is to develop a theory that explains how people learn to perform such tasks, given their perceptual, cognitive, and learning capacities. The models are a tool for developing and evaluating such a theory, in conjunction with behavioral studies and evidence concerning brain function derived neuroimaging, psychophysiological methods, and studies of impaired individuals. The modeling methodology involves endowing a model with capacities and types of knowledge that approximate what a beginning reader possesses, and providing it with similar experiences. For example, our reading models were constructed with the capacity to represent different lexical codes (orthography, phonology, semantics) as well as the capacity to learn. This

knowledge may itself have been learned and depend on other capacities (e.g., perceptual, motoric) that can be explored in other models.

In practice, model performance is greatly influenced by properties of the input and output representations. The intent is for these representations to accurately reflect children's capacities and the state of their knowledge at a particular point in development (e.g., a beginning reader); however, this ideal can only be approximated in an implemented model. Such limitations eventually show up as deviations between the performance of model and human. For example, the early model by Seidenberg and McClelland (1989) used a phonological representation that limited the model's capacity to support generalization (pronunciation of nonwords such as JINJE; Besner et al., 1990). This limitation was addressed in later models using phonological representations that incorporated additional theoretical insights (Plaut et al., 1996; Harm & Seidenberg, 1999). Every model makes simplifications about some issues in order to be able to explore other ones. Eventually the simplifications themselves become the focus of further research.

Distributed Representations

The models use distributed representations in which a particular type of information is represented by a finite set of units, with each unit participateing in many patterns (Hinton et al., 1986). For example, a model might include units that correspond to phonemes or phonetic features, each of which is activated for all the words that contain that sound. This type of representation contrasts with "localist" ones in which units correspond to higher-order entitites such as words (e.g., McClelland & Rumelhart, 1981; Grainger & Jacobs, 1994).

There is a literature debating the relative virtues of distributed vs. localist representations (see e.g., Page, 2001), but I think it is misguided. A representation is localist or distributed only in relation to other entities. For example, in McClelland and Rumelhart's interactive activation model, the representations at the letter level are localist with respect to letters (each unit corresponds to one letter) but distributed with respect to words (each word corresponds to many letters; each letter unit contributes activation to many words). Harm and Seidenberg's (2004) model used localist representations of letters, phonetic features, and semantic features, but distributed representations of the spellings, sounds, and meanings of words. The contrast is not between models employing localist versus distributed representations, because all of the above models include both. Rather, there is a contrast between models that are committed to the specific claim that there are localist representations of *words* (e.g., Coltheart et al., 2001), and models for which there are no representations at this level. In keeping with common practice, I

will use the term "localist" to refer to models with word representations, although the term is not literally accurate.

Both types of models have proved useful in past research. The choice of model depends on the state of current knowledge (often localist models are employed in early explorations of phenomena), and the question being asked (e.g., the distributed models have said more about how lexical knowledge is acquired). The use of distributed representations in our models was a theoretical choice: we think they capture basic facts about how lexical codes are represented and they are relevant to capturing various behavioral phenomena (such as consistency effects; see below). Like other aspects of the theoretical framework, the use of distributed representations is also motivated by the desire to use mechanisms that are consistent with evidence about brain function, in this case the use of large networks of simple cells to encode information.[1]

Learning

Units are linked to one another to form a network; units are activated (e.g., by presenting a letter string as input) and activation spreads to other units (e.g., units representing phonological information). The connections between units carry weights that determine how much activation is passed along. The goal is to find a set of weights that allows the model to perform the task accurately and efficiently. Learning involves adjusting the weights on the basis of experience. The reading models to date have used a learning procedure ("backpropagation") in which the output that the model produces for a word is compared to the correct, target pattern (Rumelhart et al., 1986). Small adjustments to the weights are made on the basis of the discrepancy between the two. Backpropagation is a procedure for adjusting the weights efficiently: weights that contribute a great deal to the discrepancy are adjusted more than weights that contribute less. Performance improves gradually as the weights assume values that minimize this discrepancy ("error").

Backpropagation is one of a class of "supervised" learning algorithms in which the output the model computes is compared to a target pattern (see Hinton, 1989, for a review). The use of backpropagation raises two important issues. One arises at the neurophysiological level: do neurons perform anything like the backpropagation of error? They apparently do not, but there are various proposals for how the same effects could be achieved in a biologically realistic way (see O'Reilly, 1996). So, the algorithm may be accurately capturing what the brain is doing but at a level that abstracts away from the neurophysiology. A second issue arises at the behavioral level: does human learning involve anything like comparing one's behavior to a target that fully specifies the correct response? Taken literally, backpropagation suggests that learning only takes place when a person generates a response that is corrected by an omnipresent teacher. The

algorithm is clearly an idealization, but the extent to which it deviates from normal experience, and which conclusions are affected by this idealization need to be considered carefully. One way to formulate this issue is to ask: is there a basis in the child's experience (e.g., in learning to read) for the teaching signal that automatically provided by the learning algorithm? In fact, there may be several (Harm & Seidenberg, 2004):

• a literal teacher. For tasks such as reading, for which there is explicit instruction (usually), a target is often provided by a human teacher. In fact, children typically receive more types of explicit feedback than are used in training computational models; whereas the models receive feedback about the pronunciations of words, children are explicitly taught names and sounds for letters and the pronunciations of groups of letters (such as onsets and rimes). This observation illustrates a general methodological point: many of the simplifications that implementing a model demand create a more difficult learning problem than the one confronting the child. Given these simplifications one might find it remarkable that the models approximate people's performance even as well as they do![2]

• self-generating the target. The fact that an individual can both comprehend and produce language creates the opportunity for generating one's own teaching signal. Consider learning to pronounce a word aloud. Assume the child generates a pronunciation based on the current state of his/her knowledge. The child's own output (on the production side) is also an input (on the comprehension side). If the word is pronounced correctly, then it should also produce coherent patterns if passed through the comprehension system. That is, the word will generate corresponding phonological, semantic, or even orthographic codes. All of these computations would provide a basis for deciding if the letter string had been pronounced correctly. If it is pronounced incorrectly (e.g., because the child has regularized an exception such as HAVE or PINT), then it cannot be comprehended; the failure to activate semantics would itself provide a strong error signal. This procedure is not exactly like backpropagation: it provides the correct target when the word has been pronounced correctly, but when it is mispronounced, there is only a signal that an error was made, not a specification of the correct answer. In the latter case, the child may attempt another pronunciation that succeeds, or may require an explicit teaching signal. Learning an exception does, after all, require having the correct pronunciation provided by an external source, one reason we have teachers and dictionaries. Jorm and Share (1983) described a similar "self-teaching" mechanism: the child learns by pronouncing a letter string and matching the computed output to a word that the individual has learned from using speech.

There are other ways the child may determine the correct output without explicitly being told; for example, the word may be remembered from previous exposure to the text in which it occurs, or the context may provide the target information (e.g., reading the word BEAR in a picture book about a boy's beloved toy bear).

Clearly, however, the child learner is not like backpropagation insofar as complete feedback about the correct target is not available on every learning occasion. Feedback may be partial or wholly absent; the child may know that a word was misread but not how; the child may learn from his/her own computed output, whether correct or not. Backpropagation does not capture these varied circumstances. An obvious step for future research would be to examine learning under these more variable conditions, with different types of feedback (ranging from complete target specification to no feedback) on different occasions. My experience is that models that more closely model naturalistic conditions tend to perform better and in ways that correspond more closely to people. I think it likely that merely providing fully specified targets on relatively few trials would have a disproportionately large effect on performance. It takes a few exposures to learn a word like HAVE but once the word is learned only intermittent feedback is required to retain it (Zevin & Seidenberg, 2002). Providing the correct target on every trial is not ideal; it encourages the development of word-specific knowledge, whereas what is needed is a representation of what is known in a way that supports generalization (e.g., sounding out novel words). Adding imprecision to the target pattern (e.g., not fully specifying it on every trial) may support more robust learning and better generalization.

Hidden Units

The reading models include pools of units that encode orthography, phonology, and semantics. As in the Figure 1 model, there are additional "hidden" units that mediate the computations between codes. These units allow the network to encode more complex mappings. In practice, a pattern is presented to the network (e.g., over the orthographic units), and activation spreads to the hidden layer, resulting in a pattern of activation over those units. Activation also passes to the output layer (e.g., phonology). The hidden units increase the computational power of the network, that is, the range and complexity of problems the model can solve (Rumelhart et al., 1986b). However, the hidden units also have an interesting theoretical interpretation: they are the model's basis for developing underlying representations that abstract away from surface features of the input and output codes. Thus learning in such systems is not merely the creation of associations between patterns. The hidden layer allows such generalizations to be learned; the patterns of activation over the hidden units are a reduced, intermediate code formed by this abstraction process. The existence of these intermediate units gives the networks a different character than older approaches in which behavior was construed as simple stimulus-response associations or associative chains.

Experience

The other major factor that determines the model's behavior is how it is trained, i.e., how it learns from experience. The procedure used in training the model is intended to capture basic elements of the child's experience, although like other aspects of the models, it is simplified in many respects. Like children learning to read, the models learn through exposure to many examples. In the Seidenberg and McClelland (1989) model, for example, there was a training corpus consisting of about 2900 monosyllabic words. On each training trial, a word would be selected from the corpus, the model would compute its phonological output, the output would be compared to the target, and the weights would be adjusted by the learning algorithm. This process was repeated for many words. Words were sampled such that the probability of being selected was a function of a word's frequency. Thus higher frequency words such as THE were presented more often than lower frequency words such as SINGE. The theory here is that learning the correspondences between spelling and sound involves picking up on the statistical structure of this mapping as instantiated across a large pool of words. The models pick up on this implicit structure and encode it in the weights. This contrasts with the more intuitive idea that children are learning pronunciation rules. What the model learns from exposure to one word, such as SAVE, carries over to other, partially overlapping words such as GAVE and SALE. This results from two properties of the model: (a) the use of distributed representations, and (b) the fact that the same weights are used in processing all words. Thus, changes to the weights that are beneficial for SAVE also benefit performance on GAVE and many other words. The regularities in English exist at many different grain sizes. Some occur across relatively small units (e.g., initial B is always pronounced /b/), others involve larger units such as rimes (e.g., the AVE in GAVE, SAVE, and PAVE) or complex contingencies among non-adjacent letters. The weights can also encode the atypical spelling-sound correspondences that occur in words such as HAVE and PINT. In other theories these words are treated as exceptions and handled by separate learning and processing mechanisms (e.g., rote learning; a lexicon containing a list of exceptions).

The nature of the correspondences between input and output codes varies. In alphabetic orthographies, orthography and phonology are highly correlated: letters and letter patterns represent sounds. The degree to which they are correlated (the consistency of the mapping across words) varies in alphabetic writing systems. For example, English contains more irregular correspondences than in the writing system for Serbian, in which letter-sound correspondences are almost entirely predictable. The network will pick up on these regularities to whatever extent they are present in the ensemble of training items. Thus, the same network architecture and learning procedure are thought to be involved in learning the mappings in different writing systems (Seidenberg, 1992). The same principles are assumed to apply to learning the mapping between orthography and semantics, which has a somewhat different character (Van Orden et al., 1990). For monosyllabic, monomorphemic words in English, orthography is highly predictive of

phonology but not semantics. The mapping between spelling and meaning is often said to be arbitrary but the relation is actually more complicated (Seidenberg & Gonnerman, 2000). The observation is largely true of monomorphemic words; however, morphemes are orthographic-phonological units that make systematic contributions to the meanings of many words (e.g., TEACH, TEACHER, TEACHING, etc.). Again, the model will pick up on whatever regularities exist. Learning the mapping between two correlated codes such as orthography and phonology will proceed more rapidly than learning the mapping between orthography and semantics, but the latter can be learned with sufficient experience. These differences in ease of learning play an important role in models of the development of skilled reading (Harm & Seidenberg, 2004; see below).

Again one can easily see that the model is a simplification insofar as it does not capture many elements of the child's actual experience (e.g., in a classroom). Of course, nothing prevents one from developing a model that more closely matches this experience. I've already noted that schooling includes experiences that benefit the child but are not available to existing models. However, two other issues should be noted. One is that, in the absence of a cognitive or perceptual deficit, the exact details of many aspects the child's experience may not matter a great deal. For example, we construe learning the correspondences between spelling and sound as a statistical learning problem. Because the correspondences are systematic, there is considerable redundancy: only some combinations of letters and pronunciations are permitted, and patterns are repeated across many words. Given this redundancy, individuals can converge on the same knowledge despite considerable differences in experience. For example learning the standard pronunciation of –AVE depends on exposure to words like SAVE, GAVE, and CAVE, but the exact order or relative frequencies of exposure is less important. This suggests that it may not be necessary for a model to simulate any given child's exact experience in order to capture basic facts about the learning process. The situation would be different if the goal were to simulate an individual child's performance, or perhaps the order in which words are learned averaged across many children, but that is not the grain of the phenomena to which our models are addressed.

A second issue is this: how does the experience of the child in an instructional setting correspond to what happens in a model—or more importantly, in the child's brain? Consider the situation in which a teacher provides explicit instruction about how spellings are pronounced (e.g., pointing out the fact that there is the same vowel sound in the words TRAIN, CAKE, and PLAY). Clearly we could approximate this situation in a model by providing training trials on these words. The more interesting question is how the explicit instruction of the teacher is translated into events that are realized at computational and neuronal levels. It's interesting to observe that there may be significant gaps between what a teacher thinks he/she is teaching (e.g., a rule about how a vowel is pronounced, which is an explicit type of knowledge) and what is occurring at

these other levels (which are implicit). Would a teaching method that is more closely modeled on what we think is occurring at these other levels be more effective? I don't know the answer or think it is by any means obvious. Here I would just note there are differences between learning (in the neural network model sense) and instruction (in the pedagogical sense), and that a complete theory of how children learn would explain how the effects of instruction are mediated.

Why Connectionist Models?

The modeling approach involves implementing, training, and testing a model, comparing the model's behavior to data concerning human performance, and analyzing the model's behavior, among other steps. The technical aspects of the models are daunting to the many psychologists and reading researchers who are more comfortable with an informal style of theorizing in which reading mechanisms and learning procedures are described in general rather than computational terms. Moreover, we can now study reading and its brain bases using neuroimaging techniques. Neuroimaging also plays more to cognitive psychologists' traditional strengths in experimental design and data analysis. Given these circumstances, it is important to consider why it is worth building such models at all. Several of the main reasons will be considered briefly (see Seidenberg, 1993, 2005, for further discussion).

Intuition and Beyond

Connectionist models are a source of ideas about how reading is accomplished. The approach incorporates ways of thinking about how knowledge is represented, acquired, and used that deviate in many respects from intuitive, folk-psychological accounts of cognitive phenomena. As an example, people's knowledge of words is usually assumed to be stored in a dictionary-like mental lexicon with entries for individual words. In models employing distributed representations (e.g., Figure 1), there are no lexical entries; each word is represented as a pattern of activation over sets of units encoding different codes. These models nonetheless capture phenomena previously thought to require lexical entries (e.g., frequency effects) and generate novel predictions (e.g., about effects of the consistency of spelling-sound mappings on reading aloud). This step beyond intuition is an important one. The ways we usually think about reading are closely tied to intuitions about how the process works derived from extensive personal experience. As in other areas of science, however, intuitions only provide a starting point for an investigation; often what makes a theoretical idea insightful or exciting is that it departs from intuition but nonetheless manages to provide a better account of something. Often intuitions are systematically misleading (as, for example, in the well-studied case of naïve theories of physics). The need to transcend intuition is particularly acute in the case of reading because the mechanisms we are trying to explain are largely unconscious. People are aware of the outcome

of this process—that words are understood—not the mental operations involved in achieving it. Connectionist models address the nature of underlying mechanisms at a level that intuition does not easily penetrate.

Explanatory Value

Connectionist models provide the basis for developing theories that provide a deeper explanation of behavioral phenomena such as reading. Here there are two points of comparison. One is the informal style of theorizing that dominated research in neuropsychology in the 1970s and 1980s (see Patterson, Marshall & Coltheart, 1985, for examples). The other the computational model of reading developed by Coltheart and his colleagues (Coltheart et al., 1993, 2001), which employs a different modeling methodology and thereby raises questions about the goals of the modeling enterprise.

Informal models. There are two main problems with the informal style of theorizing (Seidenberg, 1993). One is that mechanisms are often invented in response to particular behavioral phenomena and so run the risk of being little more than redescriptions of them. Our approach is different: the principles that govern connectionist models of reading are not specific to this task; they are thought to reflect more general principles that govern many aspects of language and cognition and their brain bases. In this respect the approach is consistent with the fact that reading, a cultural artifact created very recently in human history, makes use of capacities (language, vision, learning, thinking) that evolved for other purposes. The other problem with more informal approaches is that it is not always clear whether the proposed mechanisms will work in the intended ways. For example, saying that words are recognized by "accessing" their entries in the mental lexicon begs difficult questions about how the lexicon is organized and how it could be searched accurately and efficiently. Implementing a connectionist or other type of computational model requires that such concepts be stated in explicit mechanistic terms and running the model provides a way of assessing their adequacy.

Modeling as data fitting. Several types of computational models have been used in cognitive science and neuroscience; in the reading area, the main alternative is Coltheart and colleagues Dual-Route Cascade model. Coltheart et al. (2001) correctly stress that their modeling methodology is different from ours (they term their approach "Old Cognitivism"). The DRC model can be seen as part of a bottom-up, data-driven approach to modeling that has a long history in cognitive psychology; many mathematical models have this character, as well as the "information processing" models of the 1970s. These models aspire to what Chomsky (1965) termed "descriptive adequacy." Researchers conduct experiments and models are developed to "fit" the data. The main criterion for evaluating a model is the range of phenomena the model

fits. Thus Coltheart et al. (2001) emphasized the twenty-some different phenomena they simulated using a single version of DRC. Our approach is different insofar as the models are only a means to an end. The goal is a theory that explains behavior (e.g., reading) and its brain bases. The models are a tool for developing and exploring the implications of a set of hypotheses concerning the neural basis of cognition. Models are judged not only with respect to their ability to account for robust findings in a particular domain such as reading but also with respect to considerations that extend beyond a single domain. These include the extent to which the underlying computational principles apply across domains, the extent to which these principles can unify phenomena previously thought to be governed by different principles, the ability of the models to explain how behavior might arise from a neurophysiological substrate, and so on. Such models aspire to what Chomsky termed "explanatory adequacy."

Seidenberg and Plaut (in press) provide a detailed comparison of these approaches. To summarize briefly, the data-fitting approach appears to be better suited to capturing the results of individual studies, because that is the major goal of the approach. A model such as DRC thus seems satisfying because it accords with the intuition that accounting for a broad range of behavioral phenomena is always a good thing. However, when one examines DRC more closely than merely counting the number of phenomena that are simulated, problems with the approach emerge. The extent to which a model developed in this manner actually fits the data is questionable. As Seidenberg and Plaut (in press) point out, DRC exhibits a striking pattern: for almost all phenomena that were studied, the model accurately simulates the results of a single experiment (e.g., the interaction of frequency and regularity; Paap & Noel, 1991) but then produces anomalous results for other studies of the same behavioral phenomenon (e.g., Seidenberg, 1985; Taraban & McClelland, 1987). Fitting the results of one study but not others in a series is a problem; one could as well choose to report a different study and conclude that the model is inadequate. The data-fitting strategy encourages tailoring a model to reproduce the results of specific studies. This results in overfitting and a failure to generalize to other studies. This is a sign that the model does not instantiate the correct principles underlying the phenomena.

Whereas the DRC approach is data-driven, the PDP approach is more theory-driven because the models derive from a set of principles concerning neural computation and behavior. These principles are themselves motivated by computational, behavioral, and neurophysiological evidence. The models are responsive to data insofar as they need to capture patterns that reflect basic characteristic of people's behavior, particularly with regard to phenomena about which the models make different predictions. The primary goal is not to implement the model that fits the most possible data; rather, it is to use evidence provided by the model, in conjunction with other evidence (e.g., about brain organization or neurophysiology; about other types of behavior) to

converge on the correct theory of the phenomena. In fact, we could always achieve better fits to particular data sets than we have reported, but at the cost of using unmotivated "tweaks" and at the risk of overfitting. In practice there is considerable feedback between modeling, theorizing, and empirical (behavioral and, more recently, neuroimaging) research. The connectionist framework provides a set of principles and concepts out of which theories can be constructed. An implemented model instantiates some of the basic principles of the theory, for the purpose of assessing their adequacy as applied to a particular domain. At the same time, exploring a computational model typically generates new insights about underlying mechanisms and novel predictions about behavior, which can result in modifications to the theory or the general principles themselves.

Of course, this approach has its own limitations. Any given model is an imperfect instantiation of the theory on which it is based. Limitations on scope are inevitable because models become too complex to run in reasonable time or too complex to analyze and because our understanding of many phenomena is too limited. This kind of simplification and idealization is common in other areas of science, but it complicates the task of assessing a model's behavior and its theoretical implications. At some level of detail every model is necessarily false; part of the science involves determining whether the model's failures are for interesting reasons (e.g., because some aspect of the theory on which it is based is wrong) or uninteresting ones (e.g., because some phenomena are outside the scope of the model). This can be determined by experimenting with the model and comparing it to other models. Again, the limitations of a given model generate questions that inspire the next generation of research.

Establishing Causal Effects

Models provide a unique way to test causal hypotheses about the bases of normal and disordered reading. To illustrate this point, consider the issue of developmental reading impairments (dyslexia). Many hypotheses about the causes of dyslexia have been proposed: that it is secondary to impaired processing of speech (Liberman & Shankweiler, 1985); that it is secondary to deficits in the processing of visual information (e.g., Livingstone et al., 1991); that it can be caused by a learning impairment that is not specific to reading (Manis & Morrison, 1985), and there are others. The evidence for these hypotheses is largely correlational. For example, poor readers tend to be poor at spoken language tasks that involve the manipulation or comparison of phonological codes (see Blachman, 2000, for a review); similarly, some poor readers exhibit deficits on visual perception tasks such as motion detection (Eden et al., 1999; but see Sperling et al., in press), and so on for other hypotheses. These correlations are highly suggestive but it is difficult to establish a causal relationship between the hypothesized deficit and impaired reading. What is required is to show how a given type of impairment produces

specific dyslexic behaviors. For example, what is nature of the phonological deficit and why would it affect reading in specific ways? Similarly, how would a deficit in some aspect of visual processing affect learning to read and pronounce words? It is often unethical or impractical to conduct the kinds of controlled experiments that might establish more direct causal connections between deficits and behavior.

In contrast, testing causal hypothesis in a computational model is simple. Several models have focused on learning to map from orthography to phonology and the task of naming words and nonwords. Many dyslexic children (often called "phonological dyslexics") are impaired on these tasks and also on spoken-language tasks that involve the use of segmental phonological information (Snowling, 1996). The hypothesis that this pattern of poor reading derives from a phonological deficit can be tested in the following way. Take a model of normal performance and, before training has begun, introduce a phonological impairment. Harm and Seidenberg (1999) did this by introducing anomalies that affected the model's capacity to represent phonological structure. They introduced either mild or severe phonological impairments and then trained the model in the normal fashion. The purpose was to examine how learning proceeds in the presence of this "congenital anomaly." The impaired models learned more slowly but, importantly, some aspects of reading were more affected than others. With a great deal of training the models could learn the pronunciations of many words but they consistently performed poorly on the task of pronouncing novel letter strings (nonwords) such as GLORP. This behavior closely resembles that of phonological dyslexics, in whom nonword naming impairments are prominent. These children have difficulty discovering the systematic relationships between orthographic patterns and phonology (i.e., the alphabetic principle); degrading the phonological representations in the model has this effect. Studies of adults with childhood diagnoses of dyslexia (e.g., Bruck, 1998) are consistent with this picture: after many years of practice, many of the dyslexics that Bruck studied had attained considerable proficiency in reading words; however, their knowledge of phonological structure and their ability to sound out nonwords continued to be limited.

Our understanding of dyslexia is still limited, and the picture is changing rapidly, with recent research focusing on the fact that dyslexics often present with multiple impairments, not just a phonological one. The outstanding question is what kind of deficit (or deficits) could underlie these various problems (see Sperling et al., 2005, for one suggestion). My point here is only that our models provide a unique way to test hypotheses about the causes of dyslexia and how they give rise to characteristic behavioral deficits. This is possible because the models are inherently developmental: they simulate the acquisition of knowledge, which can be studied under normal and atypical conditions.

Insights from Connectionist Models

With this background in hand let us consider some of the insights to have emerged from connectionist models of reading. Several generations of models have been developed, and they continue to evolve as researchers address the inherent limitations of existing models and extend the range of phenomena they address. The two issues discussed below are central ones that any adequate theory of word reading must address.

Quasiregularity

The first issue that we addressed was how knowledge of the correspondences between the written and spoken forms of language is acquired, represented, and used. This initial focus was motivated by two considerations. First, a large body of research suggests that this knowledge plays important roles in both learning to read and skilled reading (Rayner et al., 2001). Second, learning the correspondences between the written and spoken forms of language presents an interesting computational problem, the study of which is potentially revealing about broader issues concerning learning and memory. English has an alphabetic orthography in which written symbols represent sounds. The most intuitive way to characterize the correspondences between the two is in terms of rules, a particular kind of knowledge representation. The classic evidence for rule-based knowledge is the capacity to generalize; in the reading domain, this means generating pronunciations for nonwords. Learning to read is thought to involve learning spelling-sound rules, an assumption that is widely reflected in how reading is taught. The interesting observation is that the correspondences in English are not completely consistent; there are many words (such as HAVE, GIVE, SAID, WAS, WERE, PINT, ONCE, AISLE, etc.) whose pronunciations deviate from what would be expected if the system were strictly rule-governed. In standard approaches, these words are treated as exceptions that must be learned by rote. This is the core idea underlying dual-route theories.

Note, however, that the exceptions are not arbitrary. HAVE is not pronounced “glorp;” it overlaps with many other words including HAT, HAS, and HIVE. Thus, the spelling-sound correspondences of English can be said to be rule-governed only if the rules are not obliged to apply in all cases; the system admits many forms that deviate from these central tendencies in differing degrees. Seidenberg and McClelland (1989) introduced the term “quasiregular” to describe bodies of knowledge that have this character, which include many aspects of language (e.g., inflectional and derivational morphology: Seidenberg & Gonnerman, 2000).

Connectionist networks are intrinsically well-suited to the problem of learning in quasiregular domains. A connectionist network learns to map between codes (e.g., orthography and

phonology). The weights reflect the aggregate effects of training on a large corpus of words. The weights simultaneously encode both the "rule-governed" cases and the "exceptions." Seidenberg and McClelland (1989), Plaut et al. (1996) and Harm and Seidenberg (1999) presented models that acquired spelling-sound knowledge in this manner and showed that the models could account for many phenomena associated with the task of reading letter strings aloud.

Three main aspects of this research should be noted. First, it is important to recognize how much of a departure this approach represents. Prior to the development of the connectionist framework, there was little alternative to the rules plus exceptions view. If someone had asked in 1985 what kind of lexical processing system could encode both rule-governed cases and exceptions, the question would have been treated as a non sequitur. The models challenge the deep-seated intuition that behavior is rule-governed, by demonstrating that a wholly different type of mechanism can account for the phenomena, one that is consistent with other facts about learning and its brain basis. In the connectionist framework, the characterization of language as rule-governed is taken as an informal characterization of some aspects of the underlying processing system, convenient perhaps but not accurate in detail.

Second, the approach provides an alternative way of thinking about generalization: it involves using the weights that were trained on the basis of exposure to words. Thus the weights come to encode the regularities underlying MUST, DUST, and NUT, which allows the model to correctly pronounce nonwords such as NUST the first time they are presented. Generalization had previously been thought to require rules and, indeed to provide the strongest evidence for their existence. Whether this approach will be able to account for generalization in many other domains is not known, but it invites reconsideration of the kinds of evidence standardly taken as evidence for rules.

A third point is that this approach to spelling-sound knowledge makes different predictions than the dual-route model. Our approach holds that performance is affected by the consistency of the mappings between spelling and sound. Consistency is a statistical notion, in contrast to the dual-route approach's categorical distinction between rule-governed forms and exceptions. The two theories therefore make different predictions about words such as GAVE, which are rule-governed (according to DRC) but inconsistent (according to PDP) because of irregular neighbors such as HAVE. Many studies have now replicated Glushko's (1979) original findings that spelling-sound consistency affects word and nonword pronunciation. The DRC model does not capture these phenomena correctly. According to Coltheart et al. (2001), consistency effects are mostly an artifact: many of the inconsistent words used in previous studies are actually exceptions according to DRC. They also claim that consistency effects arise from "whammies" (misanalyses of words) that occur more often in inconsistent words than rule-governed ones.

However, several studies have shown that consistency effects cannot be reduced to these factors (Jared, 2002; Cortese & Simpson, 2000). When tested on the words in these studies, DRC does not reproduce the human pattern of results.

Division of Labor

The early reading models were largely concerned with the computation of phonology. Harm and Seidenberg (2004) turned to the question of how meanings are computed. They used a variant of the Figure 1 model in which there were computations from orthography to semantics, orthography to phonology, and phonology to semantics. The model also incorporated Zorzi et al.'s (1998) use of direct connections between the outer layers (e.g., orthography and semantics; semantics and phonology; orthography and phonology).[3] Given an orthographic pattern as input, the model had to compute its meaning. The model was used to address a longstanding debate concerning the role of phonological information in silent reading. Intuitions about whether this information plays any useful role in word reading vary greatly, with plausible a priori arguments on both sides. Deriving the meanings of words directly from print seems to involve fewer steps than recoding letter strings into phonological representations and then using that information to compute meaning. On the other hand, learning the mapping from orthography to semantics may be more difficult because, as previously noted, it is more arbitrary. The pendulum has swung between "direct" and "phonologically-mediated" theories with considerable regularity over the past 100 years. The Harm and Seidenberg (2004) model offers a way to break this cycle, by treating the issue as a computational one: given the above architecture, how does the model learn to compute meanings quickly and accurately? That is, what division of labor does the model converge on, given the availability of both pathways?

The simple answer is that the model uses input from both sources for most words. The pattern that emerges over the semantics reflects the joint of effects of both pathways; what one pathway contributes depends on the capacity of the other pathway. This property contrasts with standard "race" models (Papp & Noel, 1991) in which the orthography → semantics and orthography → phonology → semantics pathways are independent, with the process that finishes first determining the access of meaning. The connectionist model performs more efficiently using both pathways than either one in isolation; thus it is not a question of which pathway wins the race, but rather how they cooperatively solve the problem. Early in training, semantic activation is largely driven by input from the orthography → phonology → semantics pathway. The phonology → semantics component was trained prior to the introduction of orthography on the view that pre-readers possess this knowledge from their use of spoken language. The orthography → phonology mapping is easy to learn because the codes are highly correlated; the orthography → semantics pathway takes longer to become established because the mapping is

more arbitrary. Over time, however, orthography → semantics begins to exert its influence, particularly for higher frequency words that get trained more often. Two main factors contribute to the development of the orthography → semantics pathway. First, it is needed to disambiguate homophones such as BEAR-BARE. Second, the pathway develops in response to the requirement to compute meanings quickly. The orthography → semantics association is more arbitrary but the pathway also involves fewer intermediate steps. If the model is given subtle pressure to respond quickly, more of the work gets taken over by orthography → semantics. Note, however, that what changes is the relative division of labor between the two pathways; there is some input from both pathways for most words.

The model is consistent with many previous assertions about word reading but also differs from them in virtue of its specific computational properties. Previous researchers have noted the tradeoffs involved in using direct orthography → semantics associations (arbitrary, but no intermediate step) vs. using phonological mediation (orthography → phonology is nonarbitrary but an additional step). The model instantiates these tradeoffs but shows that an efficient solution results if the pathways jointly determine the output, with the division of labor determined by their complementary computational properties. This account is consistent with the results of behavioral studies of homophones and pseudohomophones (e.g., Van Orden et al., 1988).

Future Directions

As I have noted throughout this chapter, the models to date are limited in scope and many basic phenomena remain to be addressed. In closing I want to mention two areas in particular.

1. Lexical semantics. The models to date have barely touched on issues concerning word meaning. This is not to minimize the significance of work such as Plaut and Booth's (2000) model of semantic priming, or the Harm and Seidenberg (2004) division of labor model. However, there is a large body of excellent research on lexical semantics that has yet to be assimilated within the computational framework. This research would include:
a. studies of the semantics of verbs (e.g., Levin, 1993). Whereas much of the research on semantic priming has focused on overlap between words at the featural level (e.g., bread-cake), verbs can be similar in meaning but participate in different sentence structures (e.g., give-donate). One concern about semantic feature representations is that they seem ad hoc, but theoretical work such as Levin's and empirical work such as McRae et al.'s (in press) provide a basis for motivated representations that can do a lot of work.
b. context effects. We continue to treat meanings as fixed entities—distributed perhaps but nonetheless unvarying. This is grossly misleading. Of course words have multiple meanings and senses, but even the meaning of a seemingly concrete word such as PIANO, which is merely the

name for a kind of keyboard instrument, varies as a function of the context in which it occurs (e.g., pushing a piano vs. playing one). It would not be difficult to implement models in which different semantic patterns are computed for words as a function of contexts that pick out different features. There is also a literature on conceptual combination suggesting how people interpret novel phrases such MOUNTAIN MAGAZINE (Gagne & Shoben, 1997; see also Clark & Clark, 1979, regarding novel uses of nouns).

c. grounding representations in perception, action, affect. There is a large body of empirical research showing that word meanings are closely tied to ("grounded in") sensory, affective, and motor experience (e.g., Glenberg & Kaschek, 2003). The extent to which grasping a ball activates the same brain circuits as the literal act remains to be determined, but there is clearly some nontrivial overlap, possibly arising from how the verb was learned in the first place. (More interesting cases involve metaphors such as "grasping an idea".) When a model includes units representing features such as "yellow" or "sour" there is an implicit assumption that this knowledge was acquired through interaction with the world. One can think of the semantic representations in the network as hidden units that mediate between perceiving, acting, and perhaps other ways of interacting with the world. Similarly, phonological representations are actually hidden units that mediate between hearing and producing sounds; orthographic representations are hidden units that mediate between seeing letters and writing them. This yields a picture like the one illustrated in Figure 2. These hidden unit representations will not have exactly the same properties as the feature-based representations used in existing models.

2. connecting model and brain. In areas such as reading, considerable evidence is accumulating concerning the brain bases of the skill. The time is ripe for a reconciliation of the computational models and this evidence. I view these approaches as complementary; each approach can inform the other and together converge on the theory of behavior and its brain bases that we all want. The models have progressed to the point where they provide strong leads for what to look for in the brain (e.g., cooperative division of labor) and generate testable hypotheses (e.g, Frost et al., 2004). At the same time, neuroimaging research is yielding evidence to which the models must be responsive. For example, there is considerable evidence about the brain bases of visual aspects of reading, including the functions of the so-called visual word form area (McCandliss et al., 2004). This aspect of reading has been sorely neglected in the computational models. On the other hand, while neuroimaging evidence has begun to identify areas that support the identification of letters despite variation in size, font, and style, computational models could contribute to understanding how this is accomplished. It also appears that processing in this region is not limited to orthographic information; it is also activated by phonological properties of words (Sandak et al., 2005). Thus it appears to function like a hidden unit representation mediating computations between visual and phonological codes, rather than as a strictly letter-based code.[4] Placing the modeling and neuroimaging in a feedback relation, in which each constrains the other, seems like a powerful approach that could yield more understanding than

either method in isolation. The goal of developing an integrated account of reading behavior and its brain bases, with computational models providing the interface between the two, seems a realistic one and likely to be the focus of considerable attention.

References

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

Blachman, B.A. (2000). Phonological awareness. In Kamil, M. L., & Mosenthal, P. B. (Eds.), *Handbook of reading research, Vol. III*.; pp. 483-502; Mahwah, NJ, US : Lawrence Erlbaum Associates, Inc.,

Bruck, M. (1998). Outcomes of adults with childhood histories of dyslexia. In C. Hulme, & R. M. Joshi (Eds.), *Reading and spelling: Development and disorders.* (pp. 179-200)Lawrence Erlbaum Associates, Publishers.

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

Clark, E. & Clark, H. (1979). When nouns surface as verbs. *Language, 55(4),* 767-811.

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

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

Cortese, M.J., & Simpson, G.B. (2000). Regularity effects in word naming: What are they? *Memory and Cognition, 28,* 1269-1276.

Eden, G.F., VanMeter, J.W. & Rumsey, J.M. (1996). Abnormal processing of visual motion in dyslexia revealed by functional brain imaging. *Nature, 382,* 66-69.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(1), 71-87.

Glushko, R.J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 674-691.

Grainger, J., & Jacobs, A.M. (1994). Orthographic processing in visual word recognition. *Psychological Review, 103,* 518-565.

Harm, M., & Seidenberg, M.S. (1999). Reading acquisition, phonology, and dyslexia: Insights from a connectionist model. *Psychological Review*, 106, 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*, 662-720.

Hinton, G. (1989). Connectionist learning procedures. *Artificial Intelligence, 40,* 185-234.

Hinton, G.E., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distribute Processing: Explorations in the Microstructure of Cognition Vol. 1.* Cambridge, MA: MIT Press.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language, 46,* 723-750.

Jorm, A.F., & Share, D.L. (1983). Phonological recoding and reading acquisition. *Applied Psycholinguistics, 4,* 103-147.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago, IL: The University of Chicago Press.

Liberman, I. Y., & Shankweiler, D. (1985). Phonology and the problems of learning to read and write. *Remedial and Special Education, 6(6),* 8-17.

Livingstone, M., Rosen, G., Drislane, F., & Galaburda, A. (1991). Physiological and anatomical evidence for a magnocellular defect in developmental dyslexia. Proc. National Academy of Sciences USA, 88, 7943-7947.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (in press). Semantic feature production norms for a large set of living and nonliving things. Special issue of *Behavioral Research Methods.*

Manis, F., & Morrison, F.J. (1985). Reading disability: A deficit in rule learning? In L.S. Siegel & F.J. Morrison (Eds.), *Cognitive Development in Atypical Children*. New York: Springer-Verlag.

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, 86,* 287-330.

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 modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23(4),* 443-512.

Paap, K., & Noel, R. W. (1991). Dual route models of print to sound: Still a good horse race. *Psychological Research, 53,* 13-24.

Patterson, K., Marshall, J.C. & Coltheart, M. (1985). *Surface Dyslexia: Cognitive and Neuropsychological Studies of Phonological Reading*. London: Lawrence Erlbaum Associates.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107(4),* 786-823.

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

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

Rumelhart, D., Hinton, G., & McClelland, J.L. (1986a). A general framework for parallel distributed processing. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing, Vol. 1.* Cambridge, MA: MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986b). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing, Vol. 1.* Cambridge, MA: MIT Press.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19,* 1-10.

Seidenberg, M.S. (1992). Beyond orthographic depth in reading: Equitable division of labor. In Ram Frost and Leonard Katz (Eds.) *Orthography, Phonology, Morphology, and Meaning,* 85-118. Oxford, England: North-Holland.

Seidenberg, M.S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228-235.

Seidenberg, M.S. (1995). Visual word recognition. In J.L. Miller and P.D. Eimas (Eds.) *Handbook of Perception & Cognition, Volume 11, Speech, Language & Communication.* San Diego: Academic Press.

Seidenberg, M. S. (2005). Connectionist models of reading. *Current Directions in Psychological Science, 14,* 238-242.

Seidenberg, M.S., & Gonnerman, L. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences 4,* 353-361.

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

Seidenberg, M. S. and Plaut, D. C. (in press). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing.* Hove, UK: Psychology Press.

Snowling, M. J. (1996). Contemporary approaches to the teaching of reading. *Journal of Child Psychology and Psychiatry, 37*(2), 139-148.

Sperling, A. J., Lu, Z.-L., Manis, F.R., & Seidenberg, M.S. (2005). Deficits in perceptual noise exclusion in developmental dyslexia. *Nature Neuroscience, 8,* 862-863.

Sperling, A.J., Lu, Z.-L., Manis, F.R., & Seidenberg, M.S. (in press). Deficits in achromatic phantom contour perception in poor readers. *Neuropsychologia*.

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

Van Orden, G. C., Johnston, J. C., & Hale, B. L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 371-386.

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

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in reading and other tasks. *Journal of Memory and Language*, *47*, 1-29.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionists dual-process model. *Journal of Experimental Psychology: Human Perception and Performance, 24(4),* 1131-1161.

Footnotes

1. Interestingly, Quiroga et al. (2005) reported that single cells in human visual cortex responded to highly specific information, such as a picture of a particular famous actress. This indicates that individual cells become highly specialized, but it does not mean that the representation of Jennifer Aniston is localist. There is no reason to think the Aniston information is represented by a single neuron rather than a network, and this network may well include many neurons that are not as highly specialized.

2. This is why I found Besner et al.'s (1990) critique of the original Seidenberg and McClelland (1989) model surprising. The model performed more poorly than people on difficult nonwords such as JINJE, but the miraculous thing, for me, was that it captured as much about word reading as it did. Still, surprise factor aside, identifying where a model's performance degrades is informative and part of a normal cycle in which the limitations of one model provide the focus for additional research thereby advancing the theoretical enterprise (Seidenberg, 2005).

3. Zorzi et al. developed a connectionist reading model that included both the orthography-hidden-phonology structure of the Seidenberg and McClelland (1989) model and an additional set of direct connections from orthography to phonology. They characterized their model as a connectionist implementation of the dual-route model, with the direct connections corresponding to a sublexical route and the hidden unit pathway corresponding to a lexical route. However, as Harm and Seidenberg (2004) noted, this characterization is not wholly accurate. The direct connections did perform well on regular words and nonwords, and poorly on exceptions, similar to the traditional "nonlexical" route. When this route is lesioned, the remaining route is not able to produce correct pronunciations for either regular or exception words. Thus, exceptions required input from both pathways to be read correctly. Unlike the standard dual-route model, damage to the nonlexical route in the Zorzi et al. model would not produce phonological dyslexia (relatively preserved word reading, impaired generalization), because the "lexical" route cannot independently read any words. The addition of direct connections between input and output layers does facilitate learning, but it does not cause the model to adopt the standard dual-route model's division of labor between lexical and nonlexical processes.

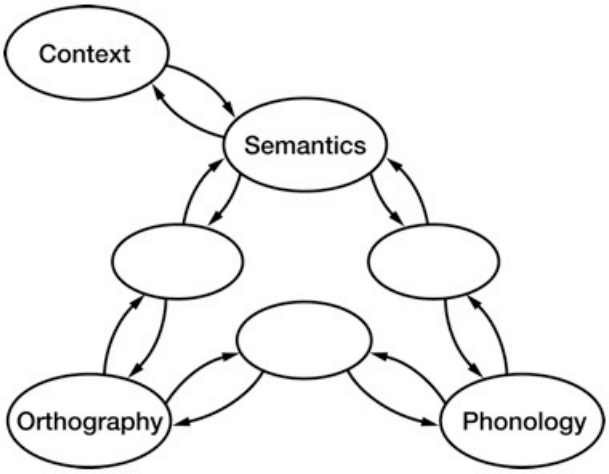
4. These findings suggest that the Harm and Seidenberg (2004) model imposes too strong a distinction between orthography and phonological processes. If orthographic representations are themselves shaped by phonology, the mapping from "orthography to

semantics" is not strictly orthographic. This would be consistent with the claim that there is partial activation of phonological information via both pathways. The division of labor issues remain essentially the same; the mapping from this representation to semantics is largely arbitrary, whereas the mapping to phonology is systematic.

Figure captions

Figure 1: The framework developed by Seidenberg and McClelland (1989). Early models focused on the orthography to phonology computation. A later model by Harm and Seidenberg (2004) addressed the computation of meaning, using a variant of this architecture.

Figure 2: Lexical codes as hidden unit representations. The semantic, phonological, and orthographic codes in existing models are simplifications; they can be seen as hidden unit representations mediating the illustrated inputs and outputs.


Context
Semantics
Orthography
Phonology


Action
Context
The world
Semantics
Hearing
Vision
Orthography
Phonology
Writing
Speaking