Connectionist Models of Word Reading

Mark S. Seidenberg

University of Wisconsin-Madison

ABSTRACT—***Connectionist models of word reading attempt to explain the computational mechanisms underlying this important skill. The goal of this research is an integrated theory of reading and its brain bases, with the computational model as the interface between the two. The models are governed by computational principles that differ considerably from naive intuitions but nonetheless account for many aspects of normal and impaired (dyslexic) reading.***

KEYWORDS: *reading; connectionist models; dyslexia*

Readers are experts at a complex, uniquely human skill, yet people's intuitions about how they read are very limited. The reading process is largely unconscious: People are aware of the outcome—comprehending a text—but not how the outcome was achieved. My theory of reading is based on connectionist models that attempt to simulate the reading process at a level that intuition does not easily penetrate. Such models serve several functions. They provide a strong test of one's theoretical assumptions: Are they sufficient to reproduce specific behavioral phenomena? One's depth of understanding can be assessed by determining whether a model instantiates general theoretical principles or has been tuned in ad hoc ways to fit specific data. Connectionist models also provide a unique way of testing causal hypotheses about reading impairments and instructional practices. For example, a hypothesis about the etiology of developmental dyslexia can be tested by configuring a model with a computational version of the impairment and seeing if the model reproduces dyslexic behavior. Finally, the models are beginning to converge with evidence about the brain bases of reading.

The main drawback of these models is that people find them difficult to understand. The technical aspects can be intimidating; the fact that they conflict with intuitions about reading doesn't help. Numerous books and Web sites provide more and less gentle introductions to technical aspects of this type of model. Here I will try to convey something about the properties of our reading models that have led to a very different understanding of this seemingly familiar skill. Note that the term "reading" covers many more phenomena than are addressed by our models, which focus on comprehending isolated words. These models represent components of a larger perceptual and cognitive system that supports text comprehension.

Address correspondence to Mark S. Seidenberg, Department of Psychology, University of Wisconsin, Madison WI 53706; e-mail: seidenberg@wisc.edu

QUASIREGULARITY

Learning the correspondences between spelling and sound is an important step in becoming a skilled reader. For years, research and teaching have been driven by the intuition that two types of knowledge are involved: Rules are used to pronounce "regular" words such as *gave* and *save*, whereas exceptions such as *have* are memorized (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Rules are also thought to support generalization, as in pronouncing a nonword such as *mave*. Dual-mechanism theories emphasize that the rule and memory subsystems are governed by different principles, acquired by different mechanisms, relevant to different types of words, and located in different brain areas. (Pinker's [1991] theory of the past tense makes similar claims.)

Assigning the rule-governed forms and exceptions to different modules creates a paradox, however, because the exceptions are not arbitrary; they overlap with the regulars. An exception such as *pint* shares structure with "rule-governed" forms such as *pant* and *pine*. Dual-mechanism theories miss these *partial regularities*. They say, in effect, that what the beginning reader learns about pronouncing *pant* and *pine* has no impact on learning *pint*, or vice versa. This seems unlikely.

Here is another possibility: The intuition that two mechanisms are necessary for learning regular words and exceptions is misleading. The system is not rule governed at all; rather it is *quasiregular*: There are different degrees of consistency in the mapping from spelling to sound. These range from rule-like (e.g., initial *b* is always pronounced /b/) to more complex contingencies. The child might learn that *-ave* is pronounced as in *gave* except in the context of *h-*, or that the *gh* in *-ght* is usually silent but not in *draught*, and so on. Many aspects of language are

quasiregular. Consider morphologically complex words: A *baker* bakes and a *thinker* thinks, but there's no corn in *corner* and a *slipper* is a kind of footwear, not a person who slips. Similarly, the past tense seems to be rule governed (*step–stepped*) but there are many partially overlapping exceptions (e.g., *sleep–slept*, *creep–crept*, *keep–kept*).

CONNECTIONIST MODELS

English spelling–sound correspondences are too complex to characterize by mere inspection. What is needed is a learning device that can discover such correspondences, to whatever degree they occur across words. Connectionist networks represent such a device and so, we think, do people.

Consider a network composed of separate groups (or "layers") of neuron-like units representing spellings (orthography) and pronunciations (phonology) of words (Fig. 1; the semantic units in the figure are discussed below). These representations are *distributed*: The finite set of units within a layer is used to represent a very large set of patterns, just as an alphabet represents many words. The orthographic representations might be composed of letters or visual features of letters; the phonological representations could be composed of phonemes (the /t/ in *bat*) or phonetic features (e.g., fricative, labial) that are the constituents of phonemes. Order of elements must also be represented: *lap* is different from *pal*. Usually there is a layer of interlevel "hidden" units that allow the network to learn and represent more complex mappings than if input and output layers were only connected directly. Processing involves activating the units corresponding to an input pattern (e.g., a word's spelling) and letting activation pass to the output units (e.g., a pronunciation) via connections between them. Each connection carries a weight that modulates the flow of activation. These elements yield a *simple feedforward network*—that is, one in which activation only flows in one direction (input → hidden → output). More complex networks are created by adding connections between the units on a layer, connections between units on the input and output layers, feedback connections (e.g., from the hidden units back to orthography), additional hidden layers, units representing the context in which a word occurs, and other computational elements.

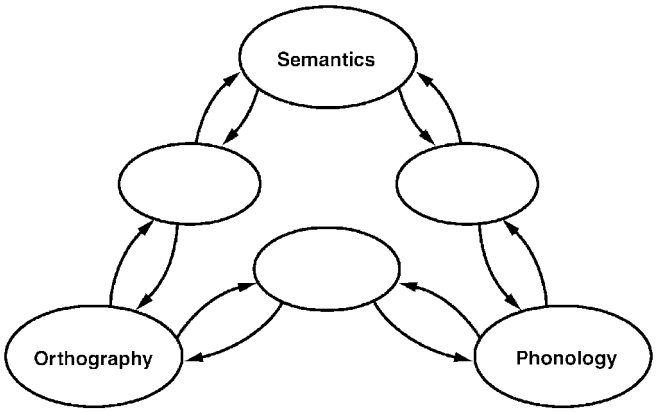

Fig. 1. Theoretical framework (introduced by Seidenberg and McClelland, 1989) that has served as the basis for several implemented models of word reading. Collectively known as the ''triangle model,'' the models differ in details and focus but are based on the same theoretical principles. Large ovals represent groups (''layers'') of units that encode different types of information: orthography (spelling), phonology (derived from pronunciation and sound), and semantics (meaning). Smaller ovals represent ''hidden units,'' which increase the computational capacity of the network and provide the basis for abstraction. Most models have focused on the orthography-to-phonology mapping. Harm and Seidenberg (2004) implemented both orthography–semantics and orthography–phonology–semantics components, using a variant of this architecture.

The model is given a task that beginning readers face: Given a spelling pattern, learn to compute its pronunciation correctly. In network terms, this means finding an appropriate set of weights. Several learning principles can be used to adjust the weights based on examples. Some principles are closely tied to how learning occurs at the neural level; some capture what is learned at a computational level that abstracts away from neurophysiological details.

Such models can learn to perform the pronunciation task accurately for thousands of words. The model represents both rule-governed cases and exceptions—*mint–pint*, *gave–have*, *bone–done*, and all the rest—contrary to the intuition that two mechanisms are necessary. How does the model do it? The pronunciation of every word involves using all the weights. All that varies on a given trial is the spelling pattern presented as input. The rest is simply a matter of computing the activations of units on successive layers. The weights have to assume values that allow all words to be pronounced correctly. This is achieved by incrementally adjusting the weights after each exposure to a word, making bigger changes to weights that contribute more to inaccurate performance.

Has the model simply memorized all of the words? No. That would require dedicating subsets of units and connections to individual words, and there aren't enough of them to achieve this. As a result, performance on any given word is affected by knowledge of other words. For example, training on *save* and *gate* results in weight adjustments that also help performance on *gave*.

According to this theory, mastering spelling–sound correspondences is a statistical learning problem. The model is a representation of this statistical knowledge; the learning algorithm is a procedure for discovering it. "Rule-governed" forms and "exceptions" represent points on a continuum of spelling–sound consistency. Many aspects of language have this graded character.

RELATING MODEL AND BEHAVIOR

Showing that a network can encode both regular and exception words is a nice trick, but what is it good for? The computational model is a simplified instantiation of a theory of behavior. If the model is based on valid principles, relevant behaviors should emerge with a minimum of fuss. For example, our pronunciation models learned to produce accurate output. However, the

weights are a compromise: What is good for *gave* is not optimal for *have* and vice versa. The net effect of the competing demands among all the words is that the model performs better on some words than on others. "Better" means either producing output that more closely matches the veridical pronunciation or producing it more rapidly (in networks that have a dynamic component and "settle into" a pronunciation over time; Plaut, McClelland, Seidenberg, & Patterson, 1996). This variability in network performance is the source of predictions about human performance: Differences in model performance should correspond to differences in human performance.

The Seidenberg and McClelland (1989) model captured several important behavioral phenomena. For example, the word-frequency effect—the observation that common words are typically read more quickly than less common words—is usually taken as evidence that words are stored as entries in lexical memory; otherwise how could people keep track of their frequencies? However, our model produced such effects even though it has no lexical entries (frequency affects the weights relevant to a word but this does not require word-level units). Moreover, it correctly predicted that frequency effects are modulated by a word's similarity to other words. The effects are smaller for words with many close neighbors (e.g., *gave*) than for "strange" words such as *sieve* or *scythe*. Frequency of exposure to *gave* is less important if the model is also acquiring friends like *gate* and *save*, whereas performance on *scythe* heavily depends on how often that word is used.

Then there are consistency effects. In a dual-mechanism theory, *must* is rule governed and *have* is an exception. But what is *gave*? *Gave* is rule governed but has the irregular neighbor *have*. Glushko (1979) found that *gave*-type words took longer to read aloud than words such as *must*, which do not have irregular neighbors. In later work, we showed that these effects are larger for lower-frequency words and less-skilled readers. Such effects are easy to explain in connectionist models. The same weights are used in pronouncing all words. Exposure to *have* shifts the weights slightly away from optimal values for *gave*, producing a small penalty such that *gave* takes longer to pronounce than words such as *must*. The effects are harder to explain in dual-route models: *Gave* and *must* are both rule governed and so should act alike. Similar effects occur for nonwords such as *mave*, both in people and in our models, a result that calls into question the fundamental idea that generalization involves applying rules.[1]

CONTROVERSIES

Connectionist models of reading have been controversial. The intuition that people learn rules and memorize exceptions is powerful and easy to grasp. The idea that the same phenomena can be explained by a multilayer network employing distributed representations and a connectionist learning algorithm is not. Moreover, it is trivially simple to falsify a computational model. Every implemented model is limited in scope, ensuring that it will fail to capture behavior at some level of detail. The question, then, is whether such limitations reflect deep flaws in the theory on which the model is based or merely the limits of a given implementation. To illustrate, our original model pronounced nonwords less well than people did (it erred on difficult ones like *faije*). If the ways that a model's performance matches people's are taken as evidence *for* the model, then surely the ways in which its performance deviates from peoples' should be taken as evidence *against* it.

Not exactly. The nonword generalization problem was soon traced to the imprecise way that phonological information was represented in the model. This imprecision had little impact on words, but affected nonwords, which require recombining known elements in novel ways. Models with improved phonological representations yielded much better nonword performance (Harm & Seidenberg, 1999; Plaut et al., 1996). Thus the nonword problem "falsified" our original model but not the theory it approximated. Moreover, this "failure" led to insights about how representations determine network behavior, to improved models, and to advances in understanding developmental dyslexia, which is associated with phonological impairments (Harm & Seidenberg, 1999). This pattern, in which the limitations of one model lead to deeper insights and improved next-generation models, is a positive aspect of the modeling methodology.

Are the models, then, unfalsifiable—able to fit any data pattern? After all, there are a lot of weights and other parameters that could be adjusted to produce particular results. In practice, this concern is moot. First, designing models that capture phenomena systematically is difficult. The model that can "fit any data" is a fiction—unfortunately! Second, tweaking a model (i.e., adjusting parameters) to fit specific behavioral results is self-defeating, because it results in *overfitting*: Matching one data set by making arbitrary changes to minor parameters will result in mismatching data from other experiments. When this occurs it is a sign the model has failed to capture relevant general principles. Finally, there is the First Law of Modeling: *Every model is false*. This property is built into the simulation methodology, because all models are limited in scope. What changes is the range of phenomena that successive models encompass and the depth of understanding of the underlying principles. So, are the models falsifiable? Positively, so to speak.

COMPUTING MEANING

Early connectionist models focused on pronouncing letter strings aloud, an interesting task and a major hurdle for beginning readers. However, the main goal in reading is under-

[1] Coltheart et al. (2001) attempted to account for consistency effects for words in terms of other factors (e.g., some of the inconsistent words used in some studies are actually exceptions according to their model). However, many studies have produced consistency effects that cannot be attributed to such factors (e.g., Jared, 2002).

standing. The technical challenges in developing models that compute meanings are substantial. Our first attempt (Harm & Seidenberg, 2004) addressed a longstanding debate: Are words read visually (computing from spelling to meaning) or phonologically (from spelling to an internal phonological code to meaning)? The pendulum has swung back and forth on this issue for many years.

Our model again departed from intuition. Previous thinking held that a meaning was accessed by either a visual or a phonological process; usually it was assumed that the two processes operated in parallel, with a "race" between them. In our model, the activation of the semantic units builds up from both pathways simultaneously. The issue is not which pathway "wins" the race but rather the division of labor between them, which varies as a function of factors such as properties of words (e.g., frequency, spelling–sound consistency) and amount of experience. Early in training, the model relied more on the orthography–phonology–semantics component (see Fig. 1); with additional training, the contribution of the orthography–semantics component increased. This model simulated various behavioral phenomena, including ones taken as evidence for other theories.

Thus, skilled reading involves the visual and phonological pathways working together. What each pathway contributes depends on what the other pathway does. The division of labor emerges as the network learns to compute meanings quickly and accurately. In skilled reading, both pathways make significant contributions to most words. The exact division of labor appears to vary between different writing systems, which differ in how they represent sound and meaning, but the computational principles are the same.

INSTRUCTION AND DYSLEXIA

What is the best way to teach reading, and what are the causes of developmental reading impairments (dyslexia)? Instructional questions can be addressed by training models in different ways; hypotheses about dyslexia can be tested by configuring models with different impairments and seeing if they develop characteristic dyslexic behaviors. The models strongly support the importance of phonics (methods that emphasize the relations between spoken and written language) in early reading instruction. In training a model, we can provide different types of feedback—for example, about both the pronunciation and meaning of a word, or just about meaning. The latter method is sometimes advocated in anti-phonics approaches. In practice, providing both types of feedback allows the model to learn and converge on an efficient division of labor more rapidly. Applications of the model to dyslexia yielded two basic findings (Harm & Seidenberg, 1999). First, the simulations supported the observation that dyslexia is often associated with impairments in the representation of phonological information. Degrading these representations causes our models to learn more slowly and to generalize poorly. Finding the neurological basis for these imprecise or "noisy" representations is a focus of current research (Sperling, Lu, Manis, & Seidenberg, 2005). Second, the models suggest that dyslexia can also have other causes. Many dyslexics exhibit a general developmental delay in reading rather than a specific phonological deficit. The modeling suggests that this delay may arise from constitutional factors (e.g., a learning deficit) or experiential ones (e.g., lack of reading experience). Some of these children may be "instructional dyslexics" who were taught using methods that did not incorporate phonics, which slows reading acquisition, as occurs in our models under similar training conditions.

FUTURE DIRECTIONS

Our reading models were developed on the basis of computational principles and behavioral phenomena. We can now study the brain bases of reading using neuroimaging. If the goal is to understand behavior in terms of the brain, where does this leave the computational models?

The simple answer is that modeling and neuroimaging methodologies are complementary. Modeling helps in understanding the brain systems that underlie complex behavior. For example, the reading brain appears to use a division-of-labor strategy like the one described above. There are two main brain circuits involved in reading: a phonologically-dominant one that develops earlier and an orthography–semantics pathway that develops with additional experience (Pugh et al., 2000). The functions of these circuits are clearer with the computational model in hand. The models also make testable neuroimaging predictions (e.g. Frost et al., 2005). At the same time, imaging research underscores the models' limitations: For example, they do not address the fact that different types of semantic information are represented in different brain regions; the representation of spelling and how it is shaped by phonological knowledge; the role of the right hemisphere in reading; the role of the hippocampus in learning; and other important issues. The reading models will continue to evolve as evidence about brain function and behavior accumulates. The goal is to converge on an integrated theory of reading behavior and its brain bases, with the computational model acting as the interface between the two. This is a powerful paradigm that can potentially be applied to many psychological phenomena.

Recommended Reading

Harm, M., & Seidenberg, M.S. (2004). (See References)

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

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

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*, *2*, 31–74.

Acknowledgments—The models described herein were developed in collaborations with James L. McClelland, David C. Plaut, and Michael Harm, whose contributions are gratefully acknowledged. My reading research is supported by National Institute of Child Health and Human Development Grant HD29891.

REFERENCES

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*, 204–256.

Frost, S.J., Mencl, W.E., Sandak, R., Moore, D.L., Rueckl, J.G., Katz, L., Fulbright, R.K., & Pugh, K.R. (2005). A functional magnetic resonance imaging study of the tradeoff between semantics and phonology in reading aloud. *NeuroReport*, *16*, 621–624.

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.

Harm, M., & Seidenberg, M.S. (1999). Reading acquisition, phonology, and dyslexia: Insights from a connectionist model. *Psychological Review*, *106*, 491–528.

Harm, M., & Seidenberg, M.S. (2004). Computing the meanings of words in reading: Division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

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

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–534.

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

Pugh, K.R., Mencl, W.E., Jenner, A.R., Katz, L., Frost, S.J., Lee, J.R., Shaywitz, S.E., & Shaywitz, B.A. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). *Mental Retardation & Developmental Disabilities Research Review*, *6*, 207–213.

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

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.