Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains

David C. Plaut
Carnegie Mellon University

James L. McClelland
Carnegie Mellon University

Mark S. Seidenberg
University of Southern California

Karalyn E. Patterson
MRC Applied Psychology Unit

Technical Report PDP.CNS.94.5
July 1994

Submitted to *Psychological Review*

Parallel Distributed Processing and Cognitive Neuroscience

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

Western Psychiatric Institute and Clinic
University of Pittsburgh
Pittsburgh, PA

Neural, Informational, and Behavioral Sciences
University of Southern California
Los Angeles, CA

MRC Applied Psychology Unit
Cambridge, England

Abstract

We develop a connectionist approach to processing in quasi-regular domains, as exemplified by English word reading. A consideration of the shortcomings of a previous implementation (Seidenberg & McClelland, 1989, *Psych. Rev.*) in reading nonwords leads to the development of orthographic and phonological representations that capture better the relevant structure among the written and spoken forms of words. In a number of simulation experiments, networks using the new representations learn to read both regular and exception words, including low-frequency exception words, and yet are still able to read pronounceable nonwords as well as skilled readers. A mathematical analysis of the effects of word frequency and spelling-sound consistency in a related but simpler system serves to clarify the close relationship of these factors in influencing naming latencies. These insights are verified in subsequent simulations, including an attractor network that reproduces the naming latency data directly in its time to settle on a response. Further analyses of the network's ability to reproduce data on impaired reading in surface dyslexia support a view of the reading system that incorporates a graded division-of-labor between semantic and phonological processes. Such a view is consistent with the more general Seidenberg and McClelland framework and has some similarities with—but also important differences from—the standard dual-route account.

Acknowledgments

This research was supported financially by the National Institute of Mental Health (Grants MH47566 and MH00385), the National Institute on Aging (Grant Ag10109), the National Science Foundation (Grant ASC–9109215), and the McDonnell-Pew Program in Cognitive Neuroscience (Grant T89–01245–016).

We thank Marlene Behrmann, Derek Besner, Max Coltheart, Joe Devlin, and Eamon Strain for helpful discussions and comments. We also acknowledge Derek Besner, Max Coltheart, and Michael McCloskey for directing attention to many of the issues addressed in this paper.

Correspondence concerning this paper should be sent to Dr. David C. Plaut, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213–3890, plaut@cmu.edu.

Contents

Introduction

Many aspects of language can be characterized as *quasi-regular*—the relationship between inputs and outputs is systematic but admits many exceptions. One such quasi-regular task is the mapping between the written and spoken forms of English words. Most words are *regular* (e.g., GAVE, MINT) in that their pronunciations adhere to standard spelling-sound correspondences. There are, however, many *exception* words (e.g., HAVE, PINT) whose pronunciations violate the standard correspondences. To make matters worse, some spelling patterns have a range of pronunciations with none clearly predominating (e.g., _OWN in DOWN, TOWN, BROWN, CROWN vs. KNOWN, SHOWN, GROWN, THROWN, or _OUGH in COUGH, ROUGH, BOUGH, THOUGH, THROUGH). Nonetheless, in the face of this complexity, skilled readers learn to pronounce written words quickly and accurately, and can also use their knowledge of spelling-sound correspondences to read pronounceable nonwords (e.g., MAVE, RINT).

An important debate within cognitive psychology is how best to characterize knowledge and processing in quasi-regular domains in order to account for human language performance. One view (e.g., Pinker, 1984, 1991) is that the systematic aspects of language are represented and processed in the form of an explicit set of rules. A rule-based approach has considerable intuitive appeal because much of human language behavior can be characterized at a broad scale in terms of rules. It also provides a straightforward account of how language knowledge can be applied productively to novel items (Fodor & Pylyshyn, 1988). However, as illustrated above, most domains are only partially systematic; accordingly, a separate mechanism is required to handle the exceptions. This distinction between a rule-based mechanism and an exception mechanism, each operating according to fundamentally different principles, forms the central tenet of so-called "dual-route" theories of language.

An alternative view comes out of research on connectionist or parallel distributed processing networks, in which computation takes the form of cooperative and competitive interactions among large numbers of simple, neuron-like processing units (McClelland, Rumelhart, & the PDP research group, 1986; Rumelhart, McClelland, & the PDP research group, 1986c). Such systems learn by adjusting weights on connections between units in a way that is sensitive to how the statistical structure of the environment influences the behavior of the network. As a result, there is no sharp dichotomy between the items that obey the rules and the items that don't. Rather, all items coexists within a single system whose representations and processing reflect the relative degree of *consistency* in the mappings for different items. The connectionist approach is particularly appropriate for capturing the rapid, online nature of language use, as well as for specifying how such processes might be learned and implemented by the brain. Perhaps more fundamentally, connectionist modeling provides a rich set of general computational principles that can lead to new and useful ways of thinking about the empirical data on human performance in quasi-regular domains.

Much of the initial debate between these two views of the language system focused on the relatively constrained domain of English inflectional morphology—specifically, forming the past-tense of verbs. Past-tense formation is a rather simple quasi-regular task: there is a single regular "rule" (add –ed; e.g., WALK ⇒"walked") and only about 100 exceptions, grouped into several clusters of similar items that undergo a similar change (e.g., SING ⇒"sang", DRINK ⇒"drank") along with a very small number of very high-frequency, arbitrary forms (e.g., GO ⇒"went"; Bybee & Slobin, 1982). Rumelhart and McClelland (1986) attempted to reformulate the issue away from a sharp dichotomy between explicit rules and exceptions, and toward a view that emphasizes the graded structure relating verbs and their inflections. They developed a connectionist model that learned to directly associate the phonology of all types of verb stems with the phonology of their past-tense forms. Pinker and Prince (1988) and Lachter and Bever (1988), however, pointed out numerous deficiencies in the model's actual performance and in some of its specific assumptions, and argued more generally that the applicability of connectionist mechanisms in language is fundamentally limited (also see Fodor & Pylyshyn, 1988). However, many of the specific limitations of the Rumelhart and McClelland model have been addressed in subsequent simulation work (Cottrell & Plunkett, 1991; Daugherty & Seidenberg, 1992; Hoeffner, 1992; MacWhinney & Leinbach, 1991; Marchman, 1993; Plunkett & Marchman, 1991, 1993). Thus, the possibility remains strong that such a model could provide a full account of past-tense inflection. Furthermore, some recent applications to aspects of language disorders (Hoeffner & McClelland, 1993; Marchman, 1993) and language change (Hare & Elman, 1992, submitted) demonstrate the ongoing extension of the approach to account for a wider range of language phenomena.

Very similar issues arise in the domain of oral reading, where there is a much richer empirical database with which to make contact. As in the domain of inflectional morphology, many researchers assume that accounting for the wealth of existing data on both normal and impaired word reading requires postulating multiple mechanisms. In particular, dual-route theorists (e.g., Besner & Smith, 1992; Coltheart, 1978, 1985; Coltheart, Curtis, Atkins, & Haller, 1993; Marshall & Newcombe, 1973; Meyer, Schvaneveldt, & Ruddy, 1974; Morton & Patterson, 1980; Paap & Noel,

1991) have claimed that pronouncing exception words requires a lexical lookup mechanism that is separate from the sublexical grapheme-phoneme correspondence (GPC) rules that apply to regular words and nonwords (also see Humphreys & Evett, 1985, and the accompanying commentaries for discussion of the properties of dual-route theories). The separation of lexical and sublexical procedures is motivated primarily by evidence that they can be independently impaired, either by abnormal reading acquisition (developmental dyslexia) or by brain damage in a previously literate adult (acquired dyslexia). Thus, *phonological* dyslexics, who can read words but not nonwords, appear to have a selective impairment of the sublexical procedure, whereas *surface* dyslexics, who can read nonwords but who "regularize" exception words (e.g., SEW ⇒"sue"), appear to have a selective impairment of the lexical procedure.

Seidenberg and McClelland (1989, hereafter SM89) challenged the central claim of dual-route theories by developing a connectionist simulation that learned to map representations of the written forms of words (orthography) to representations of their spoken forms (phonology). The network successfully pronounces both regular and exception words and yet is not an implementation of two separate mechanisms (see Seidenberg & McClelland, 1992, for a demonstration of this last point). The simulation was put forward in support of a more general framework for lexical processing in which orthographic, phonological, and semantic information interact in gradually settling on the best representations for a given input. A major strength of the approach is that it provides a natural account of the graded effects of spelling-sound consistency among words (Glushko, 1979; Jared, McRae, & Seidenberg, 1990) and how this consistency interacts with word frequency (Andrews, 1982; Seidenberg, 1985; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Furthermore, SM89 demonstrated that undertrained versions of the model exhibit some aspects of developmental surface dyslexia, and Patterson (1990, Patterson, Seidenberg, & McClelland, 1990) showed how damaging the normal model can reproduce some aspects of acquired surface dyslexia. The SM89 model also contributes to the broader enterprise of connectionist modeling of cognitive processes, in which a common set of general computational principles are being applied successfully within a wide range of cognitive domains.

However, the SM89 work has a serious empirical limitation that undermines its role in establishing a viable connectionist alternative to dual-route theories of word reading in particular, and in providing a satisfactory formulation of the nature of knowledge and processing in quasi-regular domains more generally. Specifically, the implemented model is significantly worse than skilled readers at pronouncing nonwords (Besner, Twilley, McCann, & Seergobin, 1990). This limitation has broad implications for the range of empirical phenomena that can be accounted for by the model (Coltheart et al., 1993). Poor nonword reading is exactly what would be predicted from the dual-route claim that no single system—connectionist or otherwise—can read both exception words and pronounceable nonwords adequately. Under this interpretation, the model had simply approximated a lexical look-up procedure: it could read both regular and exception words, but had not separately mastered the GPC rules necessary to read nonwords. An alternative interpretation, however, is that the empirical shortcomings of the SM89 simulation stem from specific aspects of its design and not from inherent limitations on the abilities of connectionist networks in quasi-regular domains. In particular, Seidenberg and McClelland (1990) suggested that the model's nonword reading might be improved—without adversely affecting its other properties—by using either a larger training corpus or different orthographic and phonological representations.

A second limitation of the SM89 work is that it did not provide a very extensive examination of underlying theoretical issues. SM89's main emphasis was on demonstrating that a network which operated according to fairly general connectionist principles could account for a wide range of empirical findings on normal and developmentally-impaired reading. Relatively little attention was paid in that paper to articulating the general principles themselves or to evaluating their relative importance. Thus, much of the underlying theoretical foundation of the work remained implicit. Despite subsequent efforts in explicating these principles (Seidenberg, 1993), there remains considerable confusion with regard to the role of connectionist modeling in contributing to a theory of word reading (or of any other cognitive process). Thus, some researchers (e.g., Forster, in press; McCloskey, 1991) have claimed that the SM89 demonstration, while impressive in its own right, has not extended our *understanding* of word reading because the operation of the model itself—and of connectionist networks more generally—is too complex to understand. Consequently, "connectionist networks should not be viewed as theories of human cognitive functions, or as simulations of theories, or even as demonstrations of specific theoretical points" (McCloskey, 1991, p. 387; also see Massaro, 1988; Olsen & Caramazza, 1991). Although we reject the claim that connectionist modeling is atheoretical (see Seidenberg, 1993), and that there are no bases for analyzing and understanding networks (see, e.g., Hanson & Burr, 1990), we agree that the theoretical principles and constructs for developing connectionist explanations of empirical phenomena are in need of further elaboration.

The current work develops a connectionist account of knowledge representation and cognitive processing in quasi-regular domains, in the specific context of normal and impaired word reading. The work draws on an analysis of the strengths and weaknesses of the SM89 work, with the dual aim of providing a more adequate account of the relevant empirical phenomena, and of articulating in a more explicit and formal manner the theoretical principles that underlie the approach. We explore the use of alternative representations that make the regularities between written and spoken words more explicit. In the first simulation experiment, a network using the new representations learns to read both regular and exception words, including low-frequency exception words, and yet is still able to read pronounceable nonwords as well as skilled readers. The results open up the range of possible architectures that might plausibly underlie human word reading. A mathematical analysis of the effects of word frequency and spelling-sound consistency in a related but simpler system serves to clarify the close relationship of these factors in influencing naming latencies. These insights are verified in a second simulation. Simulation 3 develops an attractor network that reproduces the naming latency data directly in its time to settle on a response, obviating the need to use error as a proxy for reaction time. The implication of the semantic contribution to reading is considered in the fourth and final simulation, in the context of accounting for the impaired reading behavior of acquired surface dyslexic patients with brain damage. Damage to the attractor network provides only a limited account of the relevant phenomena; a better account is provided by the performance of a network that learns to map orthography to phonology in the context of support from semantics. The findings lead to a view of the reading system that incorporates a graded division-of-labor between semantic and phonological processes. Such a view is consistent with the more general SM89 framework and has some similarities with—but also important differences from—the standard dual-route account. The General Discussion articulates these differences, and clarifies the implications of the current work for a broader range of empirical findings, including those raised by Coltheart et al. (1993) as challenges to the connectionist approach.

We begin with a brief critique of the SM89 model, in which we try to distinguish its central computational properties from less central aspects of its design. An analysis of its representations leads to the design of new representations that are employed in a series of simulations analogous to the SM89 simulation.

The Seidenberg and McClelland Model

The General Framework

Seidenberg and McClelland's (1989) general framework for lexical processing is shown in Figure 1. Orthographic, phonological, and semantic information is represented in terms of distributed patterns of activity over separate groups of simple neuron-like processing units. Within each domain, similar words are represented by similar patterns of activity. Lexical tasks involve transformations between these representations—for example, oral reading requires the orthographic pattern for a word to generate the appropriate phonological pattern. Such transformations are accomplished via the cooperative and competitive interactions among units, including additional *hidden* units that mediate between the orthographic, phonological, and semantic units. Unit interactions are governed by weighted connections between them, which collectively encode the system's knowledge about how the different types of information are related. The specific values of the weights are derived by an automatic learning procedure on the basis of the system's exposure to written words, spoken words, and their meanings.

The SM89 framework is broadly consistent with a more general view of information processing that has been articulated by McClelland (1991, 1993) in the context of GRAIN networks. These networks embody the following general computational principles:

- Graded: Propagation of activation is not all-or-none but rather builds up gradually over time.
- Random: Unit activations are subject to intrinsic stochastic variability.
- Adaptive: The system gradually improves its performance by adjusting weights on connections between units.
- Interactive: Information flows in a bidirectional manner between groups of units, allowing their activity levels to constrain each other and be mutually consistent.
- Nonlinear: Unit outputs are smooth, nonlinear functions of their total inputs, significantly extending the computational power of the entire network beyond that of purely linear networks.

The acronym GRAIN is also intended to convey the notion that cognitive processes are expressed at a finer grain of analysis, in terms of interacting groups of neuron-like units, than is typical of most "box-and-arrow" information

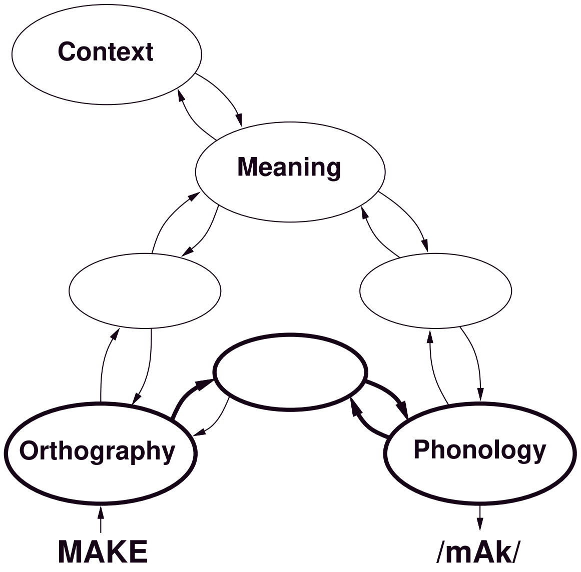

Figure 1: Seidenberg and McClelland's (1989) general framework for lexical processing. Each oval represents a group of units and each arrow represents a group of connections. The implemented model is shown in bold. (Adapted from Seidenberg & McClelland, 1989, p. 526)

processing models. Further computational principles that are central to the SM89 framework but not captured by the acronym are:

- Distributed Representations: Items of interest in the domain are represented by patterns of activity over groups of units that participate in representing many other items.

- Distributed Knowledge: Knowledge about the relationship between items is encoded across large numbers of connection weights that also encode many other mappings.

Much of the controversy surrounding the SM89 framework, and the associated implementation, stems from the fact that it breaks with traditional accounts of lexical processing (e.g., Coltheart, 1985; Morton & Patterson, 1980) in two fundamental ways. The first is in the representational status of words. Traditional accounts assume that words are represented in the structure of the reading system—in its *architecture*. Morton's (1969) "logogens" are well-known instances of this type of word representation. By contrast, within the SM89 framework the lexical status of a string of letters or phonemes is not reflected in the structure of the reading system. Rather, words are distinguished from nonwords only by *functional* properties of the system—the way in which particular orthographic, phonological, and semantic patterns of activity interact (also see Van Orden, Pennington, & Stone, 1990).

The SM89 framework's second major break with tradition concerns the degree of uniformity in the mechanism(s) by which orthographic, phonological, and semantic representations interact. Traditional accounts assume that pronouncing exception words and nonwords require separate lexical and sublexical mechanisms, respectively. By contrast, the SM89 framework employs far more homogeneous processes in oral reading. In particular, it eschews separate mechanisms for pronouncing nonwords and exception words. Rather, all of the system's knowledge of spelling-sound correspondences is brought to bear in pronouncing all types of letter strings. Conflicts among possible alternative pronunciations of a letter string are resolved, not by structurally distinct mechanisms, but by cooperative and competitive interactions based on how the letter string relates to all known words and their pronunciations. Furthermore, the semantic representation of a word participates in oral reading in exactly the same manner as do its orthographic and phonological representations, although the framework leaves open the issue of how important these semantic influences are in skilled oral reading.

A cursory inspection of Figure 1 might suggest that the SM89 framework is, in fact, a dual-route system: orthography can influence phonology either directly or via semantics. To clarify this possible source of confusion, we must be more explicit about typical assumptions in dual-route theories concerning the structure and operation of the different procedures. As described earlier, the central distinction in such theories is between lexical and sublexical procedures. The sublexical procedure applies GPC rules to produce correct pronunciations for regular words, reasonable pronunciations for nonwords, and incorrect, "regularized" pronunciations for exception words. The lexical procedure produces correct pronunciations for all words, and no response for nonwords. When the outputs of the two procedures conflict, as they do for exception words, some models (e.g., Paap & Noel, 1991) assume a "horse race" with the faster (typically lexical) procedure generating the actual response. Others (e.g., Monsell, Patterson, Graham, Hughes, & Milroy, 1992) suggest that output from the two procedures is pooled until a phonological representation sufficient to drive articulation is achieved (although the specific means by which this pooling occurs is rarely made explicit). The lexical procedure is often subdivided into a *direct* route that maps orthographic word representations directly onto phonological word representations, an *indirect* route that maps via semantics. In these formulations, the "dual-route" model is in a sense a three-route model, although researchers typically assume that the indirect, semantic route would be too slow to influence skilled word pronunciation (Coltheart, 1985; Patterson & Morton, 1985).

By contrast, the nonsemantic portion of the SM89 framework does not operate by applying GPC rules, but by the simultaneous interaction of units. It is also capable of pronouncing all types of input, including exception words, although the time it takes to do so depends on the type of input. Furthermore, the semantic portion of the framework does not operate in terms of whole-word representations, but rather in terms of interacting units, each of which participates in the processing of many words. In addition, nonwords may engage semantics to some degree, although the extent to which this occurs is likely to be minimal (see the discussion of lexical decision in the General Discussion). Thus, the structure and operation of the SM89 framework is fundamentally different from existing dual-route theories.

It may also help to clarify the relationship between the SM89 framework and approaches to word reading other than dual-route theories. The two main alternatives are lexical-analogy theories and multiple-levels theories. Lexical-analogy theories (Henderson, 1982; Marcel, 1980) dispense with the sublexical procedure, and propose that the lexical procedure can pronounce nonwords by synthesizing the pronunciations of orthographically similar words. Unfortunately, the way in which these pronunciations are generated and synthesized is rarely fully specified. Multiple-levels theories (Shallice & McCarthy, 1985; Shallice, Warrington, & McCarthy, 1983) dispense with the (direct) lexical route (or rather, incorporate it into the sublexical route) by assuming that spelling-sound correspondences are represented for segments of all sizes, ranging from single graphemes and phonemes to word bodies and entire morphemes.

In a way, the SM89 framework can be thought of as an integration and more detailed specification of lexical-analogy and multiple-level theories. The pronunciations of nonwords are generated on the basis of the combined influence of all known word pronunciations, with those most similar to the nonword having the strongest effect. In order for the system to pronounce exception words as well as nonwords, the hidden units must learn to be sensitive to spelling-sound correspondences of a range of sizes. The framework is also broadly consistent with Van Orden et al.'s (1990) proposal that orthography and phonology are strongly associated via covariant learning, although the SM89 framework incorporates direct interaction between orthography and semantics, which Van Orden and colleagues dispute.

The Implemented Model

The SM89 framework clearly represents a radical departure from widely held assumptions about lexical processing, but is it *plausible* as an account of human word reading? In the service of establishing the framework's plausibility, SM89 implemented a specific connectionist network that, they implicitly claimed, embodies the central theoretical tenets of the framework.

The network, highlighted in bold in Figure 1, contains three groups of units: 400 orthographic units, 200 hidden units, and 460 phonological units. The hidden units receive connections from all of the orthographic units and, in turn, send connections to all of the phonological units as well as back to all of the orthographic units. The network contains no semantic or context information.

Orthographic and phonological forms are represented as patterns of activity over the orthographic and phonological units, respectively. These patterns are defined in terms of context-sensitive triples of letters and phonemes (Wickelgren, 1969). It was computationally infeasible for SM89 to include a unit for each possible triple, so they used representations that require fewer units but preserve the relative similarities among patterns. In orthography, the letter triples to which each unit responds are defined by a table of 10 randomly selected letters (or a blank) in each of three positions. In the representation of a letter string, an orthographic unit is active if the string contains one of the letter triples than can be generated by sampling from each of the three positions of that unit's table. For example, GAVE would activate all orthographic units capable of generating _GA, GAV, AVE, or VE_.

Phonological representations are derived in an analogous fashion, except that a phonological unit's table entries at each position are not randomly selected phonemes, but rather all phonemes containing a particular phonemic feature (as defined by Rumelhart & McClelland, 1986). A further constraint is that the features for the first and third positions must come from the same phonetic dimension (e.g., place of articulation). Thus, each unit in phonology represents a particular ordered triple of phonemic features, termed a *Wickelfeature*. For example, the pronunciation /gAv/ would activate a phonological units representing the Wickelfeatures [*back*, *vowel*, *front*], [*stop*, *long*, *fricative*], and many others (given that /g/ has *back* and *stop* among its features, /A/ has *vowel* and *long*, and /v/ has *front* and *fricative*). On average, a word activates 81 (20.3%) of the 400 orthographic units, and 54 (11.7%) of the 460 phonological units. We will return to an analysis of the properties of these representations after summarizing the SM89 simulation results.

The weights on connections between units were initialized to small random values. The network then was repeatedly presented with the orthography of each of 2897 monosyllabic words, and trained both to generate the phonology of the word and to regenerate its orthography (see Seidenberg & McClelland, 1989, for details). During each sweep through the training set, the probability that a word was presented to the network was proportional to a logarithmic function of its frequency (Kucera & Francis, 1967). Processing a word involved setting the states of the orthographic units (as defined above), computing hidden unit states based on states of the orthographic units and the weights on connections from them, and then computing states of the phonological and orthographic units based on those of the hidden units. Back-propagation (Rumelhart, Hinton, & Williams, 1986b) was used to calculate how to adjust the weights to reduce the differences between the correct phonological and orthographic representations of the word and those generated by the network. These weight changes were accumulated during each sweep through the training set; at the end, the changes were carried out and the process was repeated.

The network was considered to have named a word correctly when the generated phonological activity was closer to the representation of the correct pronunciation of the word than to that of any pronunciation which differed from the correct one by a single phoneme. For the example GAVE ⇒/gAv/, the competing pronunciations are all those among /*Av/, /g*v/, or /gA*/, where /*/ is any phoneme. After 250 training sweeps through the corpus, amounting to about 150,000 word presentations, the network correctly named all but 77 words (97.3% correct), most of which were low-frequency exception words.

A considerable amount of empirical data on oral reading concerns the time it takes to name words of various types. A natural analogue in a model to naming latency in subjects would be the amount of computing time required to produce an output. SM89 could not use this measure because their network takes exactly the same amount of time—one update of each unit–to compute phonological output for any letter string. Instead, they approximated naming latency with a measure of the accuracy of the phonological activity produced by the network—the *phonological error score*. SM89 showed that the network's distribution of phonological error scores for various words replicates the effects of frequency and consistency in naming latencies found in a wide variety of empirical studies using the same words. Figure 2 shows particularly illustrative results in this regard. Comparing regular and exception words, the model shows the standard effects of frequency, regularity, and their interaction. Specifically, high-frequency words and regular words are fastest to name, while low-frequency exception words are disproportionately slow to name (Andrews, 1982; Seidenberg, 1985; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Furthermore, among low-frequency regular words, naming latency increases with the number of inconsistent neighbors (i.e., words in which the same body is pronounced differently; Glushko, 1979; Jared et al., 1990). Thus, regular inconsistent words like GAVE (cf. HAVE) are slower to name than regular consistent words like MUST (Taraban & McClelland, 1987), and ambiguous words like TOWN (cf. OWN) and LOVE (cf. STOVE) are slower still (Seidenberg et al., 1984).

The model also shows analogous effects of consistency in nonword naming latency. In particular, nonwords derived from regular consistent words (e.g., NUST from MUST) are faster to name than nonwords derived from exception words (e.g., MAVE from HAVE; Glushko, 1979; Taraban & McClelland, 1987). As mentioned in the Introduction, however, the model's nonword naming *accuracy* is much worse than that of skilled readers. Besner et al. (1990) reported that, on nonword lists from Glushko (1979) and McCann and Besner (1987), the model is only 59% and 51% correct, whereas skilled readers are 94% and 89% correct, respectively. Seidenberg and McClelland (1990) pointed out that the scoring criteria used for the network was more strict than that used for the subjects. We will return to the issue of scoring nonword reading performance—for the present purposes, it suffices to acknowledge that, even taking differences in scoring into account, the performance of the SM89 model on nonwords is inadequate.

The SM89 model replicates the effects of frequency and consistency in lexical decision (Waters & Seidenberg, 1985) when responses are based on *orthographic error scores*, which measure the degree to which the network succeeds at recreating the orthography of each input string. Again, however, the model is not as accurate at lexical decision under some conditions as are normal subjects (Besner et al., 1990; Fera & Besner, 1992).

Consistency also influences the ease with which word naming skills are acquired. Thus, less skilled readers—

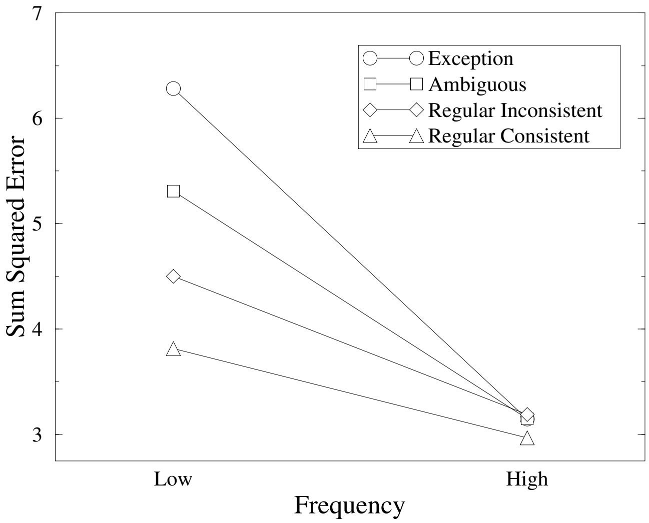

Figure 2: Mean phonological error scores produced by the Seidenberg and McClelland (1989) network for words with various degrees of spelling-sound consistency as a function of frequency. (Regenerated from Seidenberg & McClelland, 1989, p. 542, Figure 16)

whether younger or developmentally dyslexic—show larger consistency effects than more skilled readers (Backman, Bruck, Hébert, & Seidenberg, 1984; Vellutino, 1979). The model shows similar effects both early in the course of learning and when trained with limited resources (e.g., too few hidden units).

Finally, damaging the model by removing units or connections results in a pattern of errors that is somewhat similar to that of brain-injured patients with one form of surface dyslexia (Patterson, 1990; Patterson et al., 1990). Specifically, low-frequency exception words become particularly prone to being regularized (see Patterson, Coltheart, & Marshall, 1985). Overall, however, attempts to model surface dyslexia by lesioning the SM89 model have been less than completely satisfactory (see Behrmann & Bub, 1992; Coltheart et al., 1993, for criticism). We will consider this and other types of developmental and acquired dyslexia in more detail after presenting new simulation results on normal skilled reading.

Evaluation of the Model

In evaluating the SM89 results, it is important to bear in mind the relationship between the implemented model and the more general framework for lexical processing from which it was derived. In many ways, the implemented network is a poor approximation to the general framework: it contains no semantic representations or knowledge, it was trained on a limited vocabulary, and its feedforward architecture severely restricts the way in which information can interact within the system. In addition, as a working implementation, the network inevitably embodies specific representational and processing details that are not central to the overall theoretical framework. Such details include the specific orthographic and phonological representation schemes, the logarithmic frequency compression used in training, the use of error scores to model naming latencies, and the use of a supervised, error-correcting training procedure (but see Jordan & Rumelhart, 1992). Nonetheless, the implemented network is faithful to most of the central theoretical tenets of the general framework (see also Seidenberg, 1993): (a) the network employs distributed orthographic and phonological representations that reflect the similarities of words within each domain, (b) the computation of orthography and phonology involve nonlinear cooperative and competitive influences governed by weighted connections between units, (c) these weights encode all of the network's knowledge about how orthography and phonology are related, and (d) this knowledge is acquired gradually on the basis of the network's exposure to written words and their pronunciations. It is important to note that two central principles are lacking in the implemented network: interactivity and intrinsic variability. We consider the implications of these principles later.

Before we focus on the limitations of SM89's work, it is important to be clear about its strengths. First and foremost, the general framework is supported by an explicit computational model that actually implements the mapping from orthography to phonology. Certainly, implementing a model doesn't make it any more correct, but it does, among other things, allow it to be more thoroughly and adequately evaluated (Seidenberg, 1993). Many models of reading are no more explicit than “box-and-arrow” diagrams accompanied by descriptive text on how processing would occur in each component (a notable recent exception to this is the implementation of Coltheart et al., 1993, which is compared in detail with the current approach by Seidenberg, Plaut, Petersen, McClelland, & McRae, in press). In fact, the SM89 general framework amounts to such a description. By taking the further step of implementing a portion of the framework and testing it on the identical stimuli used in empirical studies, SM89 enabled the entire approach to be evaluated in much greater detail than has been possible with previous, less explicit models.

Furthermore, it should not be overlooked is that the implemented model succeeds in accounting for a considerable amount of data on normal and impaired word reading. The model reproduces the quantitative effects found in over 20 empirical studies on normal reading, as well as some basic findings on developmental and acquired dyslexia. No other existing implementation covers anything close to the same range of results.

Finally, it is important to bear in mind that the basic computational properties of the SM89 framework and implementation were not developed specifically for word reading. Rather, they derive from the much broader enterprise of connectionist modeling in cognitive domains. The same principles of distributed representations, interactivity, distributed knowledge, and gradient-descent learning are also being applied successfully to problems in high-level vision, learning and memory, speech and language, reasoning and problem solving, and motor planning and control (see Hinton, 1991; McClelland et al., 1986; Quinlan, 1991, for examples). Two distinctive aspects of the connectionist approach are its strong emphasis on general learning principles, and its attempt to make contact with neurobiological as well as cognitive phenomena. Neurally plausible learning is particularly critical to understanding reading as it is unlikely that the brain has developed innate, dedicated circuitry for such an evolutionarily recent skill. Thus, the SM89 work not only makes specific contributions to the study of reading, but also fits within a general computational approach for understanding how cognitive processes are learned and implemented in the brain.

The SM89 implementation does, however, have serious limitations in accounting for some empirical data. Some of these limitations no doubt stem from the lack of unimplemented portions of the framework—most importantly, the involvement of semantic representations, but also perhaps visual and articulatory procedures. A full consideration of the range of relevant empirical findings will be better undertaken in the General Discussion in the context of the new simulation results. Consideration of the poor nonword reading performance of the SM89 network, however, cannot be postponed. This limitation is fundamental as nonword reading is unlikely to be improved by the addition of semantics. Furthermore, Coltheart et al. (1993) have argued that, as a result of its poor nonword reading, the model is incapable of accounting for five of six central issues in normal and impaired word reading. More fundamentally, by not reading nonwords adequately, the model fails to refute the claim dual-route theorists that reading nonwords and reading exception words requires separate mechanisms.

Seidenberg and McClelland (1990) argued that the model’s poor nonwords reading was not a fundamental problem with the general framework, but rather was the result of two specific limitations in the implementation. The first is the limited size of the training corpus. The model was exposed to only about 3000 words, whereas the skilled readers with whom it is compared know approximately ten times that number. Given that the only knowledge that the model has available for reading nonwords is what it has derived from words, a limited training corpus is a serious handicap.

Coltheart et al. (1993) have argued that limitations of the SM89 training corpus cannot explain the model’s poor nonword reading because a system that learns GPC rules using the same corpus performs much better. This argument is fallacious, however, because the effectiveness of a training corpus depends critically on other assumptions built into the training procedure. In fact, Coltheart and colleagues’ procedure for learning GPC rules has built into it a considerable amount of knowledge that is specific to reading, concerning the possible relationships between graphemes and phonemes in various contexts. In contrast, SM89 applied a general learning procedure to representations that encode only ordered triples of letters and phonemic features, but nothing of their correspondences. A demonstration that the SM89 training corpus is sufficient to support good nonword reading in the context of strong, domain-specific assumptions does not invalidate the claim that the corpus may be insufficient in the context of much weaker assumptions.

The second aspect of the SM89 simulation that contributed to its poor nonword reading was the use of Wickelfeatures to represent phonology. This representational scheme has many well-known limitations, many of which are related to how well the scheme could be extended to more realistic vocabularies (see Lachter & Bever, 1988; Pinker & Prince, 1988, for detailed criticism). In the current context, Seidenberg and McClelland (1990) pointed out that the representations do not adequately capture phonemic structure. Specifically, the features of a phoneme are not bound with each other, but only with features of neighboring phonemes. As a result, the surrounding context can too easily introduce inappropriate features, producing many single-feature errors in nonword pronunciations (e.g., MAKE ⇒/nAk/ instead of /mAk/).

Neither the specific training corpus nor the Wickelfeature representation are central to the SM89 general framework for lexical processing. If Seidenberg and McClelland (1990) are correct in suggesting that it is these aspects of the simulation that are responsible for its poor nonword reading, their more general framework remains viable. On the other hand, the actual performance of their implementation is the main source of evidence that SM89 put forward in support of their view of the reading system. As McCloskey (1991) has recently pointed out, it is notoriously difficult both to determine whether a implementation's failings are due to fundamental or incidental properties of its design, and to predict how changes to its design would affect its behavior. Thus, to support the SM89 connectionist framework as a viable alternative to rule-based, dual-route accounts, it is critical to develop further simulations that account for same range of findings as the original implementation and yet also pronounce nonwords as well as skilled readers. This paper presents such simulations.

Orthographic and Phonological Representations

Wickelfeatures and the Dispersion Problem

For the purposes of supporting good nonword reading, the Wickelfeature phonological representation has a more fundamental drawback. The problem stems from the general issue of how to represent structured objects, such as words composed of ordered strings of letters and phonemes, in connectionist networks. Connectionist researchers would like their networks to have three properties (Hinton, 1990):

1. All the knowledge in a network should be in connection weights between units.

2. To support good generalization, the network's knowledge should capture the important regularities in the domain.

3. For processing to be fast, the major constituents of an item should be processed in parallel.

The problem is that these three properties are difficult to reconcile with each other.

Consider first the standard technique of using of position-specific units, sometimes called a *slot-based* representation (e.g., McClelland & Rumelhart, 1981). The first letter goes in the first slot, the second letter in the second slot, etc. Similarly for the output, the first phoneme goes in the first slot, and so on. With enough slots, words up to any desired length can be represented.

This scheme satisfies properties (1) and (3) but at a cost to property (2). That is, processing can be done in parallel across letters and phonemes using weighted connections, but at a cost of dispersing the regularities of how letters and phonemes are related. The reason is that there must be a separate copy of each letter (and phoneme) for each slot, and because the relevant knowledge is embedded in connections that are specific to these units, this knowledge must be replicated in the connections to and from each slot. To some extent this is useful in the domain of pronouncing words because spelling-sound correspondences can vary depending on whether they occur at the beginning, middle, or end of words. However, the slot-based approach carries this to an extreme, with unfortunate consequences. Consider the words LOG, GLAD, and SPLIT. The fact that the letter L corresponds to the phoneme /l/ in these words must be learned and stored three separate times in the system. There is no generalization of what is learned about letters in one position to the same letter in other positions. The problem can be alleviated to some degree by aligning the slots in various ways (e.g., centered around the vowel; Daugherty & Seidenberg, 1992) but it isn't eliminated completely (see Table 1). Adequate generalization still requires learning the regularities separately across several slots.

An alternative scheme is to apply the network to a single letter at a time, as in Sejnowski and Rosenberg's (1987) NETtalk model.[1] Here, the same knowledge is applied to pronouncing a letter regardless of where it occurs in a word, and words of arbitrary length can be processed. Unfortunately, properties (1) and (2) are now being traded off against property (3). Processing becomes slow and sequential, which may be satisfactory in many domains but not in word reading.

The representations used by SM89 were an attempt to avoid the specific limitations of the slot-based approach, but in the end turn out to have a version of the same problem. Elements such as letters and phonemes are represented, not in terms of their absolute spatial position, or relative position within the word, but in terms of the adjacent elements to the left and right. This approach, which originated with Wickelgren (1969), makes the representation of each element context sensitive without being rigidly tied to position. Unfortunately, however, the knowledge of spelling-sound correspondences is still dispersed across a large number of different contexts, and adequate generalization still requires

[1] Bullinaria (submitted) has recently developed a series of networks of this form that exhibit impressive performance in reading nonwords.

Table 1: The dispersion problem.

Slot-based representations										
Left-justified						Vowel-centered				
1	2	3	4	5		−3	−2	−1	0	1
L	O	G						S	U	N
G	L	A	D				S	W	A	M
S	P	L	I	T		S	P	L	I	T

Context-sensitive triples (“Wickelgraphs”)					
LOG:			_LO	LOG	OG_
GLAD:		_GL	GLA	LAD	AD_
SPLIT:	_SP	SPL	PLI	LIT	IT_

that the training effectively covers them all. Returning to Table 1, although the words LOG, GLAD, and SPLIT share the correspondence L ⇒/l/, they have no triples of letters in common. A similar property holds in phonology among triples of phonemes or phonemic features. Thus, as was the case in the slot-based approach, although the same correspondence is present in these three cases, different units are activated. As a result, the knowledge that is learned in one context—encoded as connection weights—does not apply in other contexts, thereby hindering generalization.

Notice that the effect of dispersing regularities is much like the effect of limiting the size of the training corpus. The contribution that an element makes to the representation of the word is specific to the context in which it occurs. As a result, the knowledge learned from one item is beneficial only to other items which share that specific context. When representations disperse the regularities in the domain, the number of trained mappings that support of given pronunciation is effectively reduced. As a result, generalization to novel stimuli, as in the pronunciation of nonwords, is based on less knowledge and suffers accordingly. In a way, Seidenberg and McClelland's (1990) two suggestions for improving their model's nonword reading performance—enlarge the training corpus and improve the representations—amount to the same thing. By using improved representations that minimize the dispersion problem, the effective size of the training corpus for a given pronunciation is increased.

Condensing Spelling-Sound Regularities

The hypothesis that guided the current work was the idea that the dispersion problem is what prevented the SM89 network from exploiting the structure of the English spelling-to-sound system as fully as human readers do. We set out, therefore, to design representations that minimize this dispersion.

The limiting case of our approach would be to have a single set of letter units, one for each letter in the alphabet, and a single set of phoneme units, one for each phoneme. Such a scheme satisfies all three of Hinton's (1990) desired properties: All of the letters in a word map to all of its phonemes simultaneously via weighted connections (and presumably hidden units), and the spelling-sound regularities are condensed because the same units and connections are involved whenever a particular letter or phoneme is present. Unfortunately, this approach has a fatal flaw: it does not preserve the relative *order* of letters and phonemes. Thus, it cannot distinguish TOP from POT or SALT from SLAT.

It turns out, however, that a scheme that involves only a small amount of replication is sufficient to uniquely represent virtually all uninflected monosyllables. By definition, a monosyllable contains only a single vowel, so only one set of vowel units is needed. A monosyllable may contain both an initial and a final consonant cluster, and almost every consonant can occur in either cluster, so separate sets of consonant units are required for each of these clusters. The remarkable thing is that this is nearly all that is necessary. The reason is that, within an initial or final consonant cluster, there are strong phonotactic constraints that arise in large part from the structure of the articulatory system. At both ends of the syllable, each phoneme can occur only once, and the order of phonemes is strongly constrained. For example, if the phonemes /s/, /t/ and /r/ all occur in the onset cluster, they must be in that order, /str/. Given this, all that is required to specify a pronunciation is which phonemes are present in each cluster—the phonotactic constraints uniquely determine the order in which these phonemes occur.

The necessary phonotactic constraints can be expressed simply by grouping phonemes into mutually exclusive sets, and ordering these sets from left to right in accordance with the left-to-right ordering constraints within consonant clusters. Once this is done, reading out a pronunciation involves simply concatenating the phonemes that are active in

Table 2: The phonological and orthographic representations.

	Phonology[a]								
onset	s S C	z Z j f v T D p b t d k g m n h	l r w y						
vowel	a e i o u @ ∧ A E I O U W Y								
coda	r	l	m n N	b g d	ps ks ts	s z	f v p k	t	S Z T D C j
	Orthography								
onset	Y S P T K Q C B D G F V J Z L M N R W H CH GH GN PH PS RH SH TH TS WH								
vowel	E I O U A Y AI AU AW AY EA EE EI EU EW EY IE OA OE OI OO OU OW OY UE UI UY								
coda	H R L M N B D G C X F V J S Z P T K Q BB CH CK DD DG FF GG GH GN KS LL NG NN PH PP PS RR SH SL SS TCH TH TS TT ZZ U E ES ED								

[a] /a/ in POT, /@/ in CAT, /e/ in BED, /i/ in HIT, /o/ in DOG, /u/ in GOOD, /A/ in MAKE, /E/ in KEEP, /I/ in BIKE, /O/ in HOPE, /U/ in BOOT, /W/ in NOW, /Y/ in BOY, /∧/ in CUP, /N/ in RING, /S/ in SHE, /C/ in CHIN /Z/ in BEIGE, /T/ in THIN, /D/ in THIS. All other phonemes are represented in the conventional way (e.g., /b/ in BAT). The groupings indicate sets of mutually exclusive phonemes.

Note: The notation for vowels is slightly different from that used by Seidenberg and McClelland (1989). Also, the representations differ slightly from those used by Plaut and McClelland (1993, Seidenberg et al., in press). In particular, /C/ and /J/ have been added for /tS/ and /dZ/, the ordering of phonemes is somewhat different, the mutually exclusive phoneme sets have been added, and the consonantal graphemes U, GU and QU have been eliminated. These changes better capture the relevant phonotactic constraints and simplify the encoding procedure for converting letter strings into activity patterns over grapheme units.

sequence from left to right, including at most one phoneme per mutually exclusive set (see Table 2).

There are a small number of cases in which two phonemes can occur in either order within a consonant cluster (e.g., /p/ and /s/ in CLASP and LAPSE). To handle such cases, it is necessary to add units to disambiguate the order (e.g., /ps/). The convention is that, if /s/ and /p/ are both active, they are taken in that order unless the /ps/ unit is active, in which case the order is reversed. To cover the pronunciations in the SM89 corpus, only three such units are required: /ps/, /ks/ and /ts/. Interestingly, these combinations are sometimes treated as single phonemes, called *affricates*, and are sometimes written with single letters (e.g., Greek ψ, English X).

This representational scheme applies almost as well to orthography as it does to phonology because English is an alphabetic language (i.e., parts of the written form of a word correspond to parts of its spoken form). However, the spelling units that correspond to phonemes are not necessarily single letters. Rather, they are what Venezky (1970) termed *relational units*, sometimes called graphemes, that can consist of from one to four letters (e.g., L, TH, TCH, EIGH). As the spelling-sound regularities of English are primarily grapheme-phoneme correspondences, the regularities in the system are most elegantly captured if the orthographic units represent the graphemes present in the string rather than simply the letters that make up the word.

Unfortunately, it is not always clear what graphemes are present in a word. Consider the word SHEPHERD. In this case, there is a P next to an H, so we might suppose that the word contains a PH grapheme, but in fact it does not; if it did it would be pronounced "she-ferd." It is apparent that the input is ambiguous in such cases. Because of this, there is no simple procedure for translating letter strings into the correct sequence of graphemes. It is, however, completely straightforward to translate a letter sequence into a pattern of activity representing all possible graphemes in the string. Thus, whenever a multiletter grapheme is present, its components are also activated. This procedure is also consistent with the treatment of affricates in phonology.

To this point, the orthographic and phonological representations have been motivated purely by computational considerations: to condense spelling-sound regularities in order to improve generalization. Before turning to the simulations, however, it is important to be clear about the empirical assumptions that are implicit in the use of these representations. Certainly, a full account of reading behavior would have to include a specification of how the representations themselves develop prior to and during the course of reading acquisition. Such a demonstration is beyond the scope of the current work. In fact, unless we are to model everything from the eye to the mouth, we cannot avoid making assumptions about the reading system's inputs and outputs, even though, in actuality, these are learned, internal representations. The best we can do is to ensure that these representations are at least broadly consistent with the relevant developmental and behavioral data.

The relevant assumptions about the phonological representations are that they are segmental (i.e., they are composed of phonemes) and that they are strongly constrained by phonotactics. We presume that this phonological structure is learned, for the most part, prior to reading acquisition, on the basis of speech comprehension and production. This is not to deny that phonological representations may become further refined over the course of reading acquisition, particularly under the influence of explicit phoneme-based instruction (see, e.g., Morais, Cary, Alegria, & Bertelson, 1979; Morais, Bertelson, Cary, & Alegria, 1986). For simplicity, however, our modeling work uses fully developed phonological representations from the outset of training.

Analogous assumptions apply with regard to the orthographic representations. We assume that they are based on letters and letter combinations, and that the ordering of these obeys orthotactic constraints (although such constraints are generally weaker than those in phonology). While these properties are not particularly controversial *per se*, orthographic representations must develop *concurrently* with reading acquisition. Thus, the use of fully-articulated orthographic representations from the outset of reading acquisition is certainly suspect.

Again, a complete account of how orthographic representations develop from more primitive visual representations is beyond the scope of the current work. Here we provide only a general characterization of such an account. We suppose that children first learn visual representations for individual letters, perhaps much like those of other visual objects. In learning to read, they are exposed to words that consist of these familiar letters in various combinations. Explicit representations gradually develop for letter combinations that occur often or have unusual consequences (see Mozer, 1990). In the context of oral reading, many of these combinations are precisely those whose pronunciations are not predicted by their components (e.g., TH, PH), corresponding to Venezky's (1970) relational units. Of course, explicit representations may develop for other, regularly-pronounced letter combinations. In the limit, the orthographic representation might contain all the letter combinations that occur in the language. Expanding our orthographic representation with multiletter units for all of these additional combinations would have little consequence because there would be little pressure for the network to learn anything about them, given that the correspondences of their components are already learned. In this way, the particular set of multiletter graphemes we employ can be viewed as an efficient simplification of a more general orthographic representation that would develop through exposure to letter combinations in words.

To be clear, we do not claim that the orthographic and phonological representations we use are fully general. Some of their idiosyncrasies stem from the fact that their design took into account specific aspects of the SM89 corpus. Nonetheless, we do claim that the principles on which the representations were derived—in particular, the use of phonotactic and orthotactic constraints to condense spelling-sound regularities—are general.

Simulation 1: Feedforward Network

The first simulation is intended to test the hypothesis that the use of representations which condensed the regularities between orthography and phonology would improve the nonword reading performance of a network trained on the SM89 corpus of monosyllabic words. Specifically, the issue is whether a single mechanism, in the form of a connectionist network, can learn to read a reasonably large corpus of words, including many exception words, and yet also read pronounceable nonwords as well as skilled readers. If such a network can be developed, it would undermined the claims of dual-route theorists that skilled word reading requires the separation of lexical and sublexical procedures for mapping print to sound.

Method

Network Architecture

The architecture of the network, shown in Figure 3, consists of three layers of units. The input layer of the network contains 105 *grapheme* units, one for each grapheme in Table 2. Similarly, the output layer contains 61 *phoneme* units. Between these two layer is an intermediate layer of 100 *hidden* units. Each unit j has a real-valued activity level or state, s_j, that ranges between 0.0 and 1.0, and is a smooth, nonlinear (logistic) function of the unit's total input, x_j.

$$x_j = \sum_i s_i w_{ij} \tag{1}$$

$$s_j = \sigma(x_j) = \frac{1}{1+\exp(-x_j)} \tag{2}$$

where w_{ij} is the weight from unit i to unit j and exp$(\cdot)$ is the exponential function.

Each hidden unit receives a connection from each grapheme unit, and in turn sends a connection to each phoneme

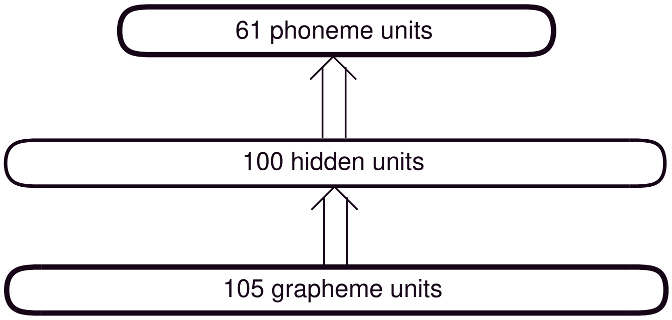

Figure 3: The architecture of the feedforward network. Ovals represent groups of units, and arrows represent complete connectivity from one group to another.

unit. In contrast to the SM89 network, the grapheme units do not receive connections back from the hidden units. Thus, the network only maps from orthography to phonology, not also from orthography to orthography (also see Phillips, Hay, & Smith, 1993). Weights on connections are initialized to small, random values, uniformly distributed between ± 0.1. Hidden and phoneme units also have associated with them a real-valued bias, which can be though of as the weight on an additional connection from a unit whose state is always 1.0 (and so can be learned in the same way as other connection weights). Including biases, the network has a total of 17,061 connections.

Training Procedure

The training corpus consists of the 2897 monosyllabic words in the SM89 corpus, augmented by 101 monosyllabic words missing from that corpus but used as word stimuli in various empirical studies, for a total of 2998 words.[2] Among these are 13 sets of homographs (e.g., READ ⇒/rEd/ and READ ⇒/red/)—for these, both pronunciations are included in the corpus. Most of the words are uninflected, although there are a few inflected forms that have been used in some empirical studies (e.g., ROLLED, DAYS). Although the orthographic and phonological representations are not intended to handle inflected monosyllables, they happen to be capable of representing those in the training corpus and so these were left in. It should be kept in mind, however, that the network's exposure to inflected forms is extremely impoverished relative to that of skilled readers.

A letter string is presented to the network by clamping the states of the grapheme units representing graphemes contained in the string to 1.0, and the states of all other grapheme units to 0.0. In processing the input, hidden units compute their states based those of the grapheme units and the weights on connections from them (according to Equations 1 and 2) and then phoneme units compute their states based on those of the hidden units. The resulting pattern of activity over the phoneme units represents the network's pronunciation of the input letter string.

After each word is processed by the network during training, back-propagation (Rumelhart et al., 1986b) is used to calculate how to change the connection weights so as to reduce the discrepancy between the pattern of phoneme activity generated by the network and the correct pattern for the word (i.e., the derivative of the error with respect to each weight). A standard measure of this discrepancy, and the one used by SM89, is the summed squared error (SSE) between the generated and correct output (phoneme) states.

$$\mathrm{SSE} = \sum_i (s_i - t_i)^2 \tag{3}$$

where s_i is the state of phoneme unit i and t_i is its correct (target) value. However, in the new representation of phonology, each unit can be interpreted as an independent hypothesis that a particular phoneme is present in the output pronunciation.[3] In this case, a more appropriate error measure is the *cross-entropy* (CE) between the generated

[2] The Plaut and McClelland (1993, Seidenberg et al., in press) network was also trained on 103 isolated grapheme-phoneme correspondences, as an approximation to the explicit instruction many children receive in learning to read. These correspondences were *not* included in the training of any of the networks reported in this paper.

[3] This is not precisely true because the procedure for determining the pronunciation based on phoneme unit activities, soon to be described, does not consider these units independently, and their states are not determined independently but are based on the same set of hidden unit states. Nonetheless, the approximation is sufficient to make cross-entropy a more appropriate error measure than summed squared error.

and correct activity patterns (see Hinton, 1989; Rumelhart, Durbin, Golden, & Chauvin, in press), also termed the assymetric divergence or Kullback-Leibler distance (Kullback & Leibler, 1951).

$$\mathrm{CE} = -\sum_i t_i \log_2(s_i) + (1 - t_t)\log_2(1 - s_i) \tag{4}$$

Notice that the contribution to cross-entropy of a given unit i is simply $-\log_2(s_i)$ if its target is 1.0, and $-\log_2(1 - s_i)$ if its target is 0.0. From a practical point of view, cross-entropy has an advantage over summed squared error when it comes to correcting output units that are completely incorrect (i.e., on the opposite flat portion of the logistic function). This is a particular concern in tasks in which output units are off for most inputs—the network can eliminate almost all of its error on the task by turning *all* of the output units off regardless of the input, including those few that should be on for this input. The problem is that, when a unit's state falls on a flat portion of the logistic function, very large weight changes are required to change its state substantially. As a unit's state diverges from its target, the change in cross-entropy increases much faster than that of summed squared error (exponentially vs. linearly) so that cross-entropy is better able to generate sufficiently large weight changes.[4]

During training, weights were also given a slight tendency to decay towards zero. This was accomplished by augmenting the cross-entropy error function with a term proportional (with a constant of 0.0001 in the current simulation) to the sum of the squares of each weight, $\sum_{i<j} w_{ij}^2$.

In the SM89 simulation, the probability that a word was presented to the network for training during an epoch was a logarithmic function of its written frequency (Kucera & Francis, 1967). In the current simulation, the same compressed frequency values are used instead to scale error derivatives calculated by back-propagation. This manipulation has essentially the same effect: more frequent words have a stronger impact than less frequent words on the knowledge learned by the system. In fact, using frequencies is this manner is exactly equivalent to updating the weights after each sweep through an expanded training corpus in which the number of times a word is present is proportional to its (compressed) frequency. The new procedure was adopted for two reasons. First, by presenting the entire training corpus every epoch, learning rates on each connection could be adapted independently (Jacobs, 1988; but see Sutton, 1992, for a recently developed on-line version).[5] Second, by implementing frequencies with multiplication rather than sampling, any range of frequencies can be used; later we will investigate the effects of using the actual Kucera and Francis (1967) frequencies in simulations. SM89 were constrained to use a logarithmic compression because less-severe compressions would have meant that the lowest frequency words might never have been presented to their network.

The actual weight changes administered at the end of an epoch are a combination of the accumulated frequency-weighted error derivatives and a proportion of the previous weight changes.

$$w_{ij}^{[t]} = \epsilon\, \epsilon_{ij} \left(\frac{\partial E}{\partial w_{ij}} + \alpha\, w_{ij}^{[t-1]} \right) \tag{5}$$

where t is the epoch number, ϵ is the global learning rate (0.001 in the current simulation), ϵ_{ij} is the connection-specific learning rate, E is the cross-entropy error function with weight decay, and α is the contribution of past weight changes, sometimes termed *momentum* (0.9 after the first 10 epochs in the current simulation). Momentum is introduced only after the first few initial epochs to avoid magnifying the effects of the initial weight gradients, which are very large because, for each word, any activity of all but a few phoneme units—those that should be active—produces a large amount of error (Plaut & Hinton, 1987).

[4] The derivative of cross-entropy with respect to an output unit's total input is simply the difference between the unit's state and its target.

$$\frac{\partial \mathrm{CE}}{\partial \mathrm{x_j}} = \frac{\partial \mathrm{CE}}{\partial \mathrm{s_j}} \frac{ds_j}{dx_j} = \left(\frac{1 - t_j}{1 - s_j} - \frac{t_j}{s_j} \right) s_j(1 - s_j) = s_j - t_j$$

[5] The procedure for adjusting the connection-specific learning rates, called delta-bar-delta (Jacobs, 1988), works as follows. Each connection's learning rate is initialized to 1.0. At the end of each epoch, the error derivative for that connection calculated by back-propagation is compared with its previous weight change. If they are both in the same direction (i.e., have the same sign), the connection's learning rate is incremented (by 0.1 in the current simulation); otherwise, it is decreased multiplicatively (by 0.9 in the current simulation).

Testing Procedure

The network, as described above, learns to take activity patterns over the grapheme units and produce corresponding activity patterns over the phoneme units. The behavior of human subjects in oral reading, however, is better described in terms of producing phoneme strings in response to letter strings. Accordingly, to directly compare the network's behavior with that of subjects, we need a procedure for encoding letter strings as activity patterns over the grapheme units, and another procedure for decoding activity patterns over the phoneme units into phoneme strings.

The encoding procedure is the one used to generate the input to the network for each word in the training corpus. To convert a letter string into an activity pattern over the grapheme units, the string is parsed into onset consonant cluster, vowel, and final (coda) consonant cluster. This involves simply locating in the string the leftmost contiguous block composed of the letters A, E, I, O, U, or (non-initial) Y. This block of letters is encoded using vowel graphemes listed in Table 2—any grapheme contained in the vowel substring is activated; all others are left inactive. The substrings to the right and left of the vowel substring are encoded similarly using the onset and coda consonant graphemes, respectively. Notice that, in a word like GUEST, the U is parsed as a vowel although it functions as a consonant (cf. GUST; Venezky, 1970). This is much like the issue with PH in SHEPHERD—such ambiguity is left for the network to cope with. The analogous encoding procedure for phonemes used to generate the training patterns for words is even simpler as monosyllabic pronunciations must contain exactly one vowel.

The decoding procedure for producing pronunciations from phoneme activities generated by the network is likewise straightforward. As shown in Table 2, phonemes are grouped into mutually exclusive sets, and these sets are ordered left to right (and top to bottom in the Table). This grouping and ordering encode the phonotactic constrains that are necessary to disambiguate pronunciations. The response of the network is simply the ordered concatenation of all active phonemes (i.e., with state above 0.5) that are the most active in their set. There are only two exceptions to this rule. The first is that, as monosyllabic pronunciations must contain a vowel, the most active vowel is included in the network's response regardless of its activity level. The second exception to relates to the affricate units, /ps/, /ks/ and /ts/. As described earlier, if an affricate is active along with its components, the order of those components in the response is reversed.

The simplicity of these encoding and decoding procedures is a significant advantage of the current representations over those use by SM89. In the latter case, reconstructing a unique string of phonemes corresponding to a pattern of activity over triples of phonemic features is exceedingly difficult, and sometimes impossible (also see Rumelhart & McClelland, 1986; Mozer, 1991). In fact, SM89 did not confront this problem—rather, they simply selected the best among a set of alternative pronunciations based on their error scores. In a sense, the SM89 model doesn't produce explicit pronunciations; it enables another procedure to select among alternatives. In contrast, the current decoding procedure does not require externally generated alternatives; every possible pattern of activity over the phoneme units corresponds directly and unambiguously to a particular string of phonemes. Nonetheless, it should be kept in mind that the encoding and decoding procedures are external to the network and, hence, constitute additional assumptions about the nature of the knowledge and processing involved in skilled reading, as discussed earlier.

Results

Word Reading

After 300 epochs of training, the network correctly pronounces all of the 2972 nonhomographic words in the training corpus. For each the 13 homographs, the network produces one of the correct pronunciations, although typically the the competing phonemes for the two alternatives are about equally active. For example, the network pronounces LEAD as /lEd/; the activation of the /E/ is 0.56 while the activation of /e/ is 0.44. These differences reflect the relative consistency of the alternatives with the pronunciations of other words.

Given the nature of the network, this level of performance on the training corpus is optimal. As the network is deterministic, it always produces the same output for a given input. Thus, in fact, it is impossible for the network to learn to produce both pronunciations of any of the homographs. Note that this determinacy is not an intrinsic limitation of connectionist networks (see, e.g., Movellan & McClelland, 1991). It merely reflects the fact that the general principle of intrinsic variability was not included in the present simulation for practical reasons—to keep the computational demands of the simulation reasonable.

For the present purposes, the important finding is that the trained network reads both regular and exception words correctly. We are also interested in how well the network replicates the effects of frequency and consistency on naming latency. However, we will return to this issue after we consider the more pressing issue of the network's performance

Table 3: Percent of "regular" pronunciations of nonwords.

	Glushko (1979)		McCann and Besner (1987)
	Regular	Exception	Control
	Nonwords	Nonwords	Nonwords
Subjects	93.8	78.3	88.6
Network	97.7	72.1	85.0

Table 4: Errors by the feedforward network in pronouncing nonwords from Glushko (1979) and McCann and Besner (1987).

Glushko (1979)			McCann and Besner (1987)		
Nonword	Correct	Response	Nonord	Correct	Response
Regular Nonwords (1/43)			Control Nonwords (12/80)		
MUNE	/myUn/	/m(y 0.43)Un/	*PHOYCE	/fYs/	/(f 0.42)Y(s 0.00)/
Exception Nonwords (12/43)			*TOLPH	/tolf/	/tOl(f 0.12)/
BILD	/bild/	/bIld/	*ZUPE	/zUp/	/zyUp/
BOST	/bost/	/bOst/	SNOCKS	/snaks/	/snask(ks 0.31)/
COSE	/kOz/	/kOs/	LOKES	/lOks/	/lOsk(ks 0.02)/
GROOK	/grUk/	/gruk/	*YOWND	/yWnd/	/(y 0.47)and/
LOME	/lOm/	/l∧m/	KOWT	/kWt/	/kOt/
MONE	/mOn/	/m∧n/	FAIJE	/fAj/	/fA(j 0.00)/
PILD	/pild/	/pIld/	*ZUTE	/zUt/	/zyUt/
PLOVE	/plOv/	/pl∧v/	*VEEZE	/vEz/	/(v 0.40)Ez/
POOT	/pUt/	/put/	*PRAX	/pr@ks/	/pr@sk(ks 0.33)/
SOOD	/sUd/	/sud/	JINJE	/jinj/	/jIn(j 0.00)/
SOST	/sost/	/s∧st/			
WEAD	/wEd/	/wed/			

Note: The activity levels of correct but missing phonemes are listed in parentheses. In these cases, the actual response is what falls outside the parentheses. Words marked with "*" remain errors after considering properties of the training corpus (as explained in the text).

in reading nonwords.

Nonword Reading

We tested the network on three lists of nonwords from two empirical studies. The first two lists comes from an experiment by Glushko (1979), in which he compared subjects' reading of 43 nonwords derived from regular words (e.g., HEAN from DEAN) with their reading of 43 nonwords derived from exception words (e.g., HEAF from DEAF). The other list come from a study by McCann and Besner (1987), in which they compared performance on a set of 80 pseudohomophones (e.g., BRANE) with a set of 80 control nonwords (e.g., FRANE). We used only their control nonwords in the present investigation as we believe pseudohomophone effects are mediated by aspects of the reading system, such as semantics, that are not implemented in our simulation (see the General Discussion).

As nonwords are, by definition, novel stimuli, exactly what counts as the "correct" pronunciation of a nonword is a matter of considerable debate (see, e.g., Masterson, 1985; Seidenberg et al., in press). The complexity of this issue will become apparent momentarily. For the purposes of an initial comparison, we will consider the pronunciation of a nonword to be correct if it is regular, as defined by adhering to the GPC rules outlined by Venezky (1970).

Table 3 presents the correct performance of skilled readers reported by Glushko (1979) and by McCann and Besner (1987) on their nonword lists, and the corresponding performance of the network. Table 4 lists the errors made by the network on these lists.

First consider Glushko's regular nonwords. The network makes only a single minor mistake on these items, just failing to introduce the transitional /y/ in MUNE. In fact, this inclusion varies across dialects of English (e.g., DUNE ⇒/dUn/ vs. /dyUn/). In the training corpus, the four words ending in _UNE (DUNE, JUNE, PRUNE, TUNE) are all coded without the /y/. In any case, overall both the network and subject have no difficult on these relatively easy nonwords.

The situation is rather different for the exception nonwords. Both the network and subjects produce non-regular pronunciations for a significant subset of these items, with the network being slightly more prone to do so. However, a closer examination of the responses in these cases reveals why. Consider the nonword GROOK. The grapheme OO is most frequently corresponds to /U/, as in BOOT, and so the correct (regular) pronunciation of GROOK is /grUk/. However, the body _OOK is almost always pronounced /u/, as in TOOK. The only exception to this among the 12 words ending in _OOK in the training corpus is SPOOK ⇒/spUk/. This suggests that /gruk/ should be the correct pronunciation.

Actually, the issue of whether the network's pronunciation is correct or not is less relevant than the issue of whether the network behaves similarly to subjects. In fact, both the subjects and the network are sensitive to the context in which vowels occur, as evidenced by their much greater tendency to produce non-regular pronunciations for exception nonwords as compared with regular nonwords. Glushko (1979) found that 80% of subject's non-regular responses to exception nonwords were consistent with some other pronunciation of the nonword's body that occurs in the Kucera and Francis (1967) corpus, leaving only 4.1% of all responses as actual errors. In the network, *all* of the non-regular responses to exception nonwords match some other pronunciation in the training corpus for the same body, with half of these being the most frequent pronunciation of the body. None of the network's responses to exception nonwords are actual errors. Overall, the network performs as well if not slightly better than subjects on the Glushko nonword lists. Appendix 1 lists all of the pronunciations accepted as correct for each of the Glushko nonwords.

Both the subjects and the network find McCann and Besner (1987) control nonwords more difficult pronounce, which is not surprising as the lists contain a number of orthographically unusual nonwords (e.g., JINJE, VAWX). Overall, the network's performance is slightly worse than that of subjects. However, many of the network's errors can be understood in terms of specific properties of the training corpus and network design. First, although there is no word in the training corpus with the body _OWT, medial OW is often pronounced /O/ (e.g., BOWL ⇒/bOl/) and so KOWT ⇒/kOt/ should be considered a reasonable response. Second, two of the errors are on inflected forms, SNOCKS and LOKES, and as previously acknowledged, the network is not intended to apply to inflections and has minimal experience with them. Finally, there are no instances in the training corpus of words containing the grapheme J in the coda, and so the network cannot possibly have learned to map it to /j/ in phonology. In a way, for a nonword like JINJE, the effective input to the network is JINE, to which the network's response /jIn/ is correct. This also applies to the nonword FAIJE. Excluding these and the inflected forms from the scoring, and considering KOWT ⇒/kOt/ correct, the network performs correctly on 69/76 (90.8%) of the remaining control nonwords, which is slightly better than the subjects. Most of the remaining errors of the network involve correspondences that are infrequent or variable in the training corpus (e.g., PH ⇒/f/, U ⇒/yU/).

It must be acknowledged that the failure of the model on inflected forms and on those with final J are real shortcomings that would have to be addressed in a completely adequate account of word reading. Our purpose in separating out these items in the above analysis simply acknowledges that the model's limitations are easily understood in terms of specific properties of the training corpus.

Is it a Dual-Route Model?

One possibility, consistent with dual-route theories, is that the network has partitioned itself into two sub-networks, one that reads regular words, and another that reads exception words. If this were the case, some hidden units would contribute to exception words but not to nonwords, while others would contribute to nonwords but not to exception words. To test this possibility, we measured the contribution a hidden unit makes to pronouncing a letter string by how much the cross-entropy error in pronouncing the string increases when the unit is removed from the network. If the network had partitioned itself, there would be a negative correlation across hidden units between the number of exception words and the number of nonwords to which each hidden unit makes a substantial contribution (greater than 0.2). In fact, for orthographically matched exception words and nonwords (Taraban & McClelland, 1987), there is a moderate *positive* correlation between the numbers of exception words and nonwords to which hidden units contribute ($r = .25$, $t_{98} = 2.59$, $p = .011$; see Figure 4). Thus, some units are more important for the overall task and some are less important, but the network has not partitioned itself into one system that learns the rules and another system that learns the exceptions.

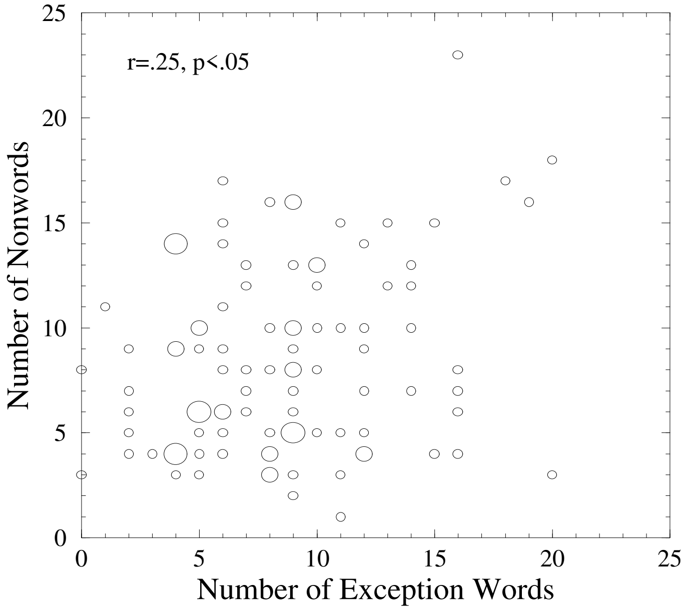

Figure 4: The numbers of exception words and nonwords ($n = 48$ for each) to which each hidden unit makes a significant contribution, as indicated by an increase in cross-entropy error of at least 0.2 when the unit is removed from the network. Each circle represents one or more hidden units, with the size of the circle proportional to the number of hidden units making significant contributions to the indicated numbers of exception words and nonwords.

Frequency and Consistency Effects

It is important to verify that, in addition to producing good nonword reading, the new model replicates the basic effects of effects of frequency and consistency in naming latency. Like the SM89 network, the current network takes the same amount of time to compute the pronunciation of any letter string. Hence, we must also resort to using an error score as an analogue of naming latency. In particular, we will use the cross entropy between the network's generated pronunciation of a word and it's correct pronunciation, as this is the measure that the network was trained to minimize. Later we will examine the effects of frequency and consistency directly in the settling time of an equivalently trained recurrent network when pronouncing various types of words.

Figure 5 shows the mean cross entropy error of the network in pronouncing words of varying degrees of spelling-sound consistency (Taraban & McClelland, 1987) as a function of frequency. Overall, high-frequency words produce less error than low-frequency words ($F_{1,184}$=17.1, p<.001). However, frequency interacts significantly with consistency ($F_{3,184}$=5.65, p=.001). Post-hoc comparisons within each word type separately reveal that the effect of frequency reaches significance at the 0.05 level only for exception words (although the effect for regular inconsistent words is significant at 0.053). The effect of frequency among all regular words (consistent and inconsistent) just fails to reach significance ($F_{1,94}$=3.14, p=.08).

There is also a main effect of consistency in the error made by the network in pronouncing words ($F_{3,184}$=24.1, p<.001). Furthermore, collapsed across frequency, all post-hoc pairwise comparisons of word types are significant. Specifically, regular consistent words produce less error than regular inconsistent words, which in turn produce less error than ambiguous words, which in turn produce less error than exception words. Interestingly, the effect of consistency is significant considering only high-frequency words ($F_{3,92}$=12.3, p<.001). All pairwise comparisons are also significant except between exception words and ambiguous words. This contrasts with the performance of normal subjects, who typically show little or no effect of consistency among high frequency words (e.g., Seidenberg, 1985; Seidenberg et al., 1984).

Summary

A feedforward connectionist network was trained on an extended version of the SM89 corpus of monosyllabic words, using orthographic and phonological representations that condense the regularities between these domains.

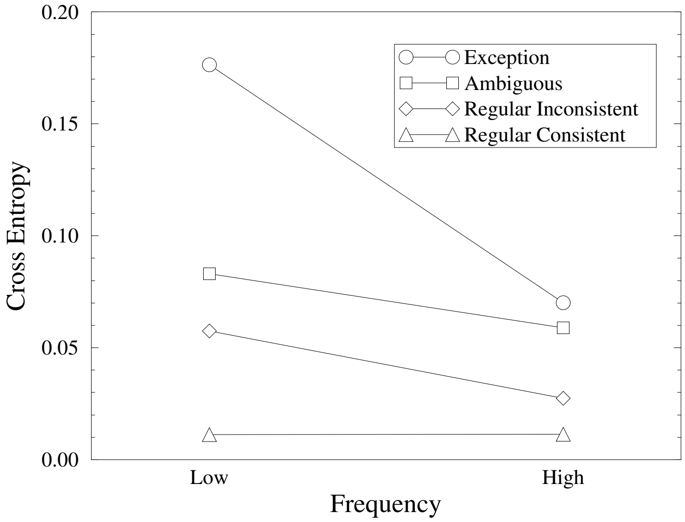

Figure 5: Mean cross-entropy error produced by the feedforward network for words with various degrees of spelling-sound consistency as a function of frequency.

After training, the network reads regular and exception words flawlessly and yet also reads pronounceable nonwords (Glushko, 1979; McCann & Besner, 1987) essentially as well as skilled readers. Minor discrepancies in performance can be ascribed to nonessential aspects of the simulation. Critically, the network had not segregated itself over the course of training into separate mechanisms for pronouncing exception words and nonwords. Thus, the network directly refutes the claims of dual-route theorists that skilled word reading requires the separation of lexical and sublexical procedures for mapping print to sound.

Furthermore, the error produced by the network on various types of words, as measured by the cross entropy between the generated and correct pronunciations, replicates the standard findings of frequency, consistency, and their interaction in the naming latencies of subjects (Andrews, 1982; Seidenberg, 1985; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). A notable exception, however, is that, unlike subjects and the SM89 network, the current network exhibits a significant effect of consistency among high-frequency words.

Analytic Account of Frequency and Consistency Effects

The main pattern of the effects of frequency and consistency on naming latency have often been interpreted as requiring explicit lexical representations and grapheme-phoneme correspondence rules. However, the SM89 network and the one presented in the previous section exhibit these effects without these properties. What, then, gives rise to this pattern of frequency and consistency effects in these networks?

The relevant empirical pattern of results can be described in the following way. In general, high-frequency words are named faster than low-frequency words, and words with greater spelling-sound consistency are named faster than words with less consistency. However, the effect of frequency diminishes as consistency is increased, and the effect of consistency diminishes as frequency is increased. A natural interpretation of this pattern is that frequency and consistency contribute independently to naming latency, but that the system as a whole is subject to what might be termed a gradual ceiling effect: the magnitude of increments in performance decreases as performance improves. Thus, if either the frequency or the consistency of a set of words is sufficiently high on its own to produce fast naming latencies, increasing the other factor will yield little further improvement.

A close analysis of the operation of connectionist networks reveals that these effects are a direct consequence of properties of the processing and learning in these networks—specifically, the principles of Nonlinearity, Adaptivity, and Distributed Representations and Knowledge referred to earlier. In a connectionist network, the weight changes induced by a word during training serve to reduce the error on that word (and hence, by definition, its naming latency). The frequency of a word is reflected in how often it is presented to the network (or, as in the previous simulation, in the explicit scaling of the weight changes it induces). Thus, word frequency directly amplifies weight changes that are helpful to the word itself.

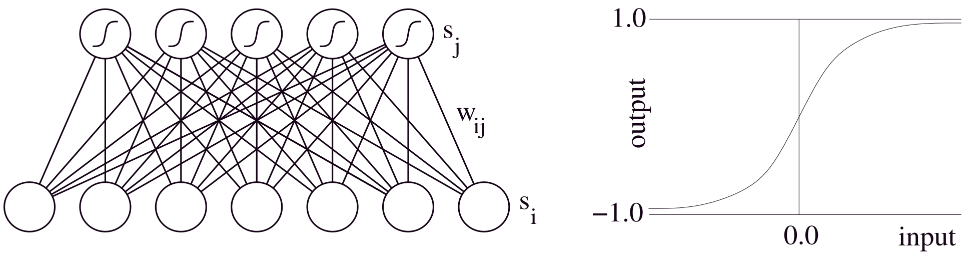

Figure 6: A simple network for analyzing frequency and consistency effects.

The consistency of the spelling-sound correspondences of two words is reflected in the similarity of the orthographic and phonological units that they activate. Furthermore, two words will induce similar weight changes to the extent that they activate similar units. Given that the weight changes induced by a word are superimposed on the weight changes for all other words, a word will tend to be helped by the weight changes for words whose spelling-sound correspondences are consistent with its own (and, conversely, hindered by the weight changes for inconsistent words). Thus, frequency and consistency effects contribute independently to naming latency because they both arise from similar weight changes that are simply added together during training.

Over the course of training, the magnitude of the weights in the network increase in proportion to the accumulated weight changes. These weight changes result in corresponding increases in the summed input to output units that should be active, and decreases in the summed input to units that should be inactive. However, due to the nonlinearity of the input-output function of units, these changes do not translate directly into proportional reductions in error. Rather, as the magnitude of the summed inputs to output units increases, their states gradually asymptote towards 0.0 or 1.0. As a result, a given increase in the summed input to a unit yields progressively smaller decrements in error over the course of training. Thus, although frequency and consistency each contribute to the weights, and hence to the summed input to units, their effect on error is subjected to a gradual ceiling effect as unit states are driven towards extremal values.

The Frequency-Consistency Equation

To see the effects of frequency and consistency in connectionist networks more directly, it will help to consider a network that embodies some of the same general principles as the SM89 and feedforward networks, but which is simple enough to permit a closed-form analysis (following Anderson, Silverstein, Ritz, & Jones, 1977, also see Stone, 1986). In particular, consider a nonlinear network without hidden units and trained with a correlational (Hebbian) rather than error-correcting learning rule (see Figure 6). Such a network is a specific instantiation of Van Orden et al.'s (1990) *covariant learning hypothesis*. To simplify the presentation, we will assume that input patterns are composed of 1's and 0's, output patterns are specified in terms of $+1$'s and -1's, and connection weights are all initialized to 0.0. We will derive an equation that expresses in concise form the effects of frequency and consistency in this network on its response to any given input.

A learning trial involves setting the states of the input units to the input pattern (e.g., orthography) for a word, setting the output units to the desired output pattern (e.g., phonology) for the word, and adjusting the weight from each input unit to each output unit according to

$$\triangle\, w_{ij} = \epsilon s_i s_j \tag{6}$$

where ϵ is a learning rate constant, s_i is the state of input unit i, s_j is the state of output unit j, and w_{ij} is the weight on the connection between them. After each input-output training pattern is presented once in this manner, the value of each connection weight is simply the sum of the weight changes for each individual pattern:

$$w_{ij} = \epsilon \sum_p s_i^{[p]} s_j^{[p]} \tag{7}$$

where p indexes individual training patterns.

After training, the network's performance on a given test pattern is determined by setting the states of the input units to the appropriate input pattern and having the network compute the states of the output units. In this computation, the state of each output unit is assumed to be a nonlinear, monotonically increasing function of the sum, over input units, of the state of the input unit times the weight on the connection from it:

$$s_j^{[t]} = \sigma \left(\sum_i s_i^{[t]} w_{ij} \right) \tag{8}$$

where t is the test pattern and $\sigma(\cdot)$ is the nonlinear input-unit function. An example of such a function, the standard logistic function commonly used in connectionist networks, is shown on the right of Figure 6. The input-output function of the output units need not be this particular function, but it must have certain of its properties: it must vary monotonically with input, and it must approach its extremal values (here, ± 1) at a diminishing rate as the magnitude of the summed input increases (positively or negatively). We call such functions *sigmoid* functions.

We can substitute the derived expression for each weight w_{ij} from Equation 7 into Equation 8, and pull the constant term ϵ out of the summation over i to obtain

$$s_j^{[t]} = \sigma \left(\epsilon \sum_i s_i^{[t]} \sum_p s_i^{[p]} s_j^{[p]} \right) \tag{9}$$

This equation indicates that the activation of each output unit reflects a sigmoid function of the learning rate constant ϵ times of a sum of terms, each consisting of the activation of one of the input units in the test pattern times the sum, over all training patterns, of the activation of the input unit times the activation of the output unit. In our present formulation, where the input unit's activation is 1 or 0, this sum reflects the extent to which the output unit's activation tends to be equal to 1 when the input unit's activation is equal to 1. Specifically, it will be exactly equal to the number of times the output unit is equal to 1 when the input unit is equal to 1, minus the number of times the output unit is equal to -1 when the input unit is equal to 1. We can see from Equation 9 that if, over an entire ensemble of training patterns, there is a consistent value of the activation of an output unit when an input unit is active, then the connection weights between them will come to reflect this. If the training patterns come from a completely regular environment, such that each output's activation depends on only one input unit and is completely uncorrelated with the activation of every other input unit, then all the weights to each output unit will equal 0 except the weight from the particular input unit on which it depends. (If the set of training patterns are sampled randomly from a larger space of patterns, the sample will not reflect the true correlations exactly, but will be scattered approximately normally around the true value.) Thus, the learning procedure discovers which output units depend on which input units, and sets the weights accordingly. For our purposes in understanding quasi-regular domains, in which the dependencies are not so discrete in character, the weights will come to reflect the degree of consistency between each input unit and each output unit, over the entire ensemble of training patterns.

Equation 9 can be written a different way to reflect a relationship that is particularly relevant to the word reading literature, in which the frequency of a particular word and the consistency of its pronunciation with the pronunciations of other, similar words are known to influence the accuracy and latency of pronunciation. The rearrangement expresses a very revealing relationship between the output at test and the similarity of the test pattern to each input pattern:

$$s_j^{[t]} = \sigma \left(\epsilon \sum_p s_j^{[p]} \sum_i s_i^{[p]} s_i^{[t]} \right) \tag{10}$$

This expression shows the relationship between the state of an output unit at test as a function of its states during training and the *similarity* between the test input pattern and each training input pattern, measured in terms of their dot product, $\sum_i s_i^{[p]} s_i^{[t]}$. For input patterns consisting of 1's and 0's, this measure amounts to the number of 1's the two patterns have in common, which we refer to as the *overlap* of training pattern p and test pattern t and designate $\mathcal{O}^{[pt]}$. Substituting into the previous expression, we find that the state of an output unit at test reflects the sum over all training patterns of the unit's output for that pattern times the overlap of the pattern with the test pattern.

$$s_j^{[t]} = \sigma \left(\epsilon \sum_p s_j^{[p]} \mathcal{O}^{[pt]} \right) \tag{11}$$

Notice that the product $s_j^{[p]} \mathcal{O}^{[pt]}$ is a measure of the input-output *consistency* of the training and test patterns. To see this, suppose that the inputs for the training and testing patterns have considerable overlap. Then the contribution of the training pattern depends on the sign of the output unit's state for that pattern. If this sign agrees with that of the appropriate state for the test pattern (i.e., the two patterns are consistent) the training pattern will help to move the state of the output unit towards the appropriate extremal value for the training pattern. However, if the signs of the states for the training and test patterns disagree (i.e., the patterns are inconsistent), performance on the test pattern is worse for having learned the training pattern. As the input for the training pattern becomes less similar to that of the test pattern, reducing $\mathcal{O}^{[pt]}$, the impact of their consistency on test performance diminishes.

To clarify the implications of the above equation, it will help to consider some simple cases. First, suppose that the network is trained on only one pattern, and tested with a variety of patterns. Then the state of each output unit during testing will be a monotonic function of its value in the training pattern times the overlap of the training and test input patterns. As long as there is any overlap in these patterns, the test output will have the same sign as the training output, and its magnitude will increase with the overlap between the test pattern and training pattern. Thus, the response of each output unit varies with the similarity of the test pattern to the pattern used in training.

As a second example, suppose we test only on the training pattern itself, but vary the number of training trials on the pattern. In this case, the summation over the p training patterns in the above equation reduces to a count of the number of training presentations of the pattern. Thus, the state of the output unit on this pattern will approach its correct asymptotic value of ± 1 as the number of training presentations increases.

Finally, consider the more general case in which several different input-output patterns are presented during training, with each one presented some number of times. Then, elaborating Equation 11, the state of an output unit at test can be written as

$$s_j^{[t]} = \sigma\left(\epsilon \sum_p F^{[p]} s_j^{[p]} \mathcal{O}^{[pt]}\right) \tag{12}$$

where $F^{[p]}$ is the number (frequency) of training presentations of pattern p.

We will refer to Equation 12 as the frequency-consistency equation. Relating this equation to word and nonword reading simply involves identifying the input to the network with a representation of the spelling of a word, and the output of the network with a representation of its pronunciation. Given the assumption that stronger activations correspond to faster naming latencies, we can use the frequency-consistency equation to derive predictions about the relative naming latencies of different types of words. In particular, the equation provides a basis for understanding why naming latency depends on the frequency of a word ($F^{[p]}$) and the consistency of its spelling-sound correspondences with those of other words ($s_j^{[p]}\mathcal{O}^{[pt]}$). It also accounts for the fact that the effect of consistency diminishes as the frequency of the word increases (and vice versa), since high-frequency words push the value of the sum out into the tail of the input-output function, where influences of other factors are reduced (see Figure 7).

Quantitative Results with a Simple Corpus

To make the implications of the frequency-consistency equation more concrete, suppose a given output unit should have a value of $+1$ if a word's pronunciation contains the vowel /I/ (as in DIVE) and -1 if it contains the vowel /i/ (as in GIVE). Suppose further that we have trained the network on a set of words ending in _IVE which all contain either /I/ or /i/ as the vowel. Then the frequency-consistency equation tells us immediately that the response to a given test input should reflect the influence of every one of these words to some degree. Holding all else constant, the higher the frequency of the word, the more closely the output will approach the desired value. Holding the frequency of the word itself constant, the more other similar words agree with its pronunciation, the more closely the output will approach the correct extremal value. The distance from the desired value will vary continuously with the difference between the total influence of the neighbors that agree with the word and the neighbors that disagree, with the contribution of each neighbor weighted by its similarity to the word and its frequency. When the word itself has a high frequency, it will tend to push the activation close to the correct extreme. Near the extremes, the slope of the function relating the summed input to the state of the output unit becomes relatively shallow, so the influence of the neighbors is diminished.

To illustrate these effects, Figure 8 shows the cross-entropy error for a particular output unit as we vary the frequency of the word being tested and its consistency with 10 other, overlapping words (also see Van Orden, 1987). For simplicity, we assume that all ten words have a frequency of 1.0 and an overlap of 0.75 with the test word—this would be true, for example, if input units represented letters and words differed in a single letter out of four. Four degrees of consistency are examined: (a) exception words (e.g., GIVE), for which all but one of the ten neighbors disagree with the test word on the value of the output unit; (b) ambiguous words (e.g., PLOW), for which the neighbors

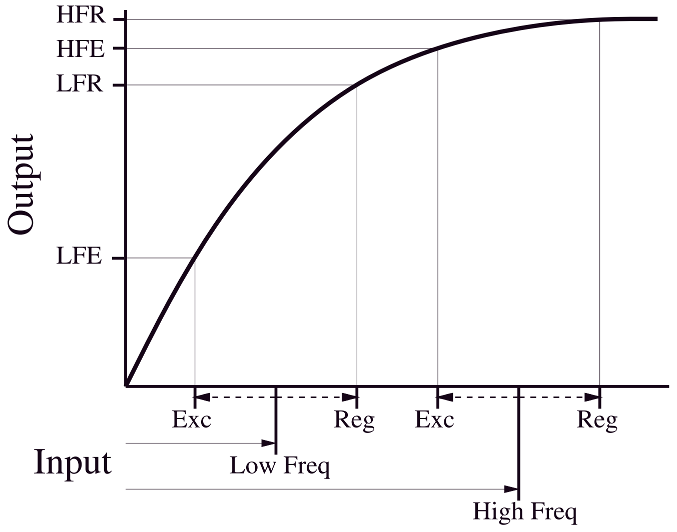

Figure 7: A frequency-by-consistency interaction arising out of applying an asymptoting output activation function to the additive input contributions of frequency (solid arrows) and consistency (dashed arrows). Notice in particular that the identical contribution from consistency has a much weaker effect on high-frequency words than on low-frequency words. Only the top half of the logistic activation function is shown.

are split evenly between those that agree and those that disagree; (c) regular inconsistent words (e.g., DIVE), for which most neighbors agree but two disagree (namely GIVE and LIVE); and (d) regular consistent words (e.g., DUST), for which all neighbors agree on value of the output unit. In the present analysis, these different cases are completely characterized in terms of a single variable: the consistency of the pronunciation of the vowel in the test word with its pronunciation in other words with overlapping spellings. The analysis clearly reveals a graded effect of consistency that diminishes with increasing frequency.

Error Correction and Hidden Units

It should be noted that the Hebbian approach described here does not, in fact, provide an adequate mechanism for learning the spelling-sound correspondences in English. For this, we require networks with hidden units trained using an error-correcting learning rule such as back-propagation. In this section we take some steps in the direction of extending the analyses to these more complex cases.

First we consider the implications of using an error-correcting learning rule rather than Hebbian learning, still within a network with no hidden units. Back-propagation is a generalization of one such rule, known as the *delta rule* (Widrow & Hoff, 1960). The first observation is that, when using the delta rule, the change in weight w_{ij} due to training on pattern p is proportional to the state of the input unit, $s_i^{[p]}$, times the partial derivative of the error on pattern p with respect to the summed input to the output unit j, $\delta_j^{[p]}$, rather than simply times the correct state of unit j, $s_j^{[p]}$ (cf. Equation 6). As a result, Equation 12 becomes

$$s_j^{[t]} = \sigma\left(\epsilon \sum_p F^{[p]} \delta_j^{[p]} \mathcal{O}^{[pt]}\right) \tag{13}$$

Matters are more complex here because $\delta_j^{[p]}$ depends on the actual performance of the network on each trial. However, $\delta_j^{[p]}$ will always have the same sign as $s_j^{[p]}$, because an output unit's error always has the same sign as its target as long as the target is an extremal value of the activation function (± 1 here), and because only unit j is affected by a change to its input. Thus, as in the Hebbian case, training on a word that is consistent with the test word will always help unit i to be correct, and training on an inconsistent word will always hurt, thereby giving rise to the consistency effect.

The main difference between the Hebb rule and the delta rule is that, with the latter, if a set of weights exists

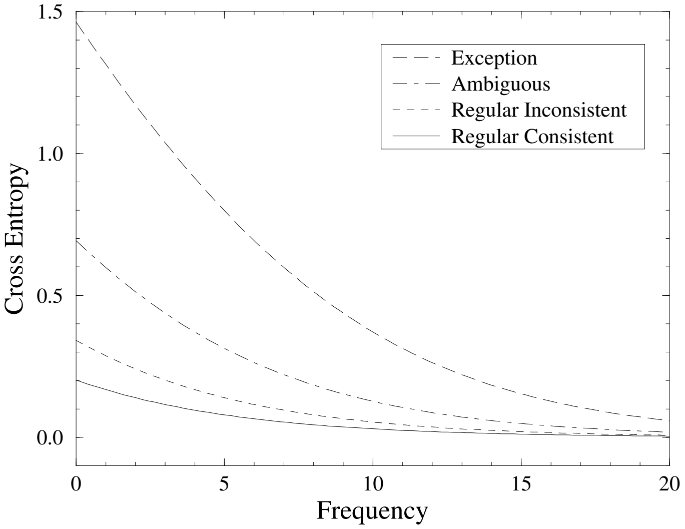

Figure 8: The effects of frequency and consistency in a network without hidden units trained with correlational (Hebbian) learning ($\epsilon = 0.2$ in Equation 12).

that allows the network to produce the correct output for each training pattern, the learning procedure will eventually converge to it.[6] This is generally not the case with Hebbian learning, which often results in responses for some cases that are incorrect. To illustrate this, we consider applying the two learning rules to a training set for which a solution does exist. The solution is found by the delta rule and not by the Hebb rule.

The problem is posed within the framework we have already been examining. The specific network consists of 11 input units (with values of 0 and 1) representing letters of a word. The input units send direct connections to a single output unit that should be $+1$ if the pronunciation of the word contains the vowel /I/ but -1 if it contains the vowel /i/. Table 5 shows the input patterns and the target output for each case, as well as the net inputs and activations that result from training with each learning rule. There are 10 items in the training set, six with the body _INT and four with the body _INE. The _INE words all take the vowel /I/, so for these the vowel has a target activation of $+1$; five of the _INT words take /i/, so the vowel has a target of -1. The _INT words also include the exception word PINT that takes the vowel /I/. For this analysis, each word is given an equal frequency of 1.

Table 6 lists the weights from each input unit to the output unit that are acquired after training with each learning rule. For the Hebb rule, this involved 5 epochs of training using a learning rate $\epsilon = 0.1$. The resulting weights are equal to 0.5 (the number of training epochs times the learning rate, ϵ) times the number of training items in which the letter is present and the vowel is /I/, minus the number of items in which the letter is present and the vowel is /i/. Specifically, the letters L and M occur once with /I/ and once with /i/, so their weight is 0.0; the letters I and N occur five times with /I/ and five time with /i/, so their weights are also 0.0. Final E and final T have the largest magnitude weights; E is strongly positive because it occurs four times with /I/ and never with /i/, and T is strongly negative because it occurs five times with /i/ and only once with /I/. A and F are weakly positive since each occurs once with /I/, and D, H and T are weakly negative since each occurs once with /i/. P is moderately positive, since it occurs twice with /I/—once in PINE and once in PINT. Thus, these weights directly reflect the co-occurrences of letters and phonemes.

The outputs produced by the network when using the weights produced by the Hebb rule, shown in Table 5, illustrate the consistency effect, both in net inputs and in activations. For example, the net input for FINE is stronger than for LINE, because LINE is more similar to the inconsistent LINT; and the net input for PINE is stronger than for LINE, since PINE benefits from its similarity with PINT, which has the same correspondence. However, the weights do not completely solve the task: For the word PINT, the net input is -1.0 (1.0 from the P minus 2.0 from the T), and passing this through the logistic function results in an activation of -0.46, which is quite different from the target value of $+1$.

[6] Actually, given the use of extremal targets and an asymptoting activation function, no set of finite weights will reduce the error to zero. In this case, a "solution" consists of a set of weights that produces outputs that are closer than some specified tolerance (say, 0.01) to the target value for every output unit in every training pattern. If a solution exists that produces outputs that all have the correct sign (tolerance of 1.0, given targets of ± 1), then a solution also exists for any smaller tolerance because multiplying all the weights by a large enough constant will push the output of the sigmoid arbitrarily close to its extreme values without affecting its sign.

Table 5: Input Patterns, Targets, and Activations after Training with Hebb Rule and Delta Rule

	Letter Inputs												Hebb Rule		Delta Rule	
Word	D	F	H	L	M	P	T	I	N	E	T	Target	Net	Act	Net	Act
DINT	1	0	0	0	0	0	0	1	1	0	1	−1	−2.5	−0.85	−2.35	−0.82
HINT	0	0	1	0	0	0	0	1	1	0	1	−1	−2.5	−0.85	−2.29	−0.82
LINT	0	0	0	1	0	0	0	1	1	0	1	−1	−2.0	−0.76	−1.70	−0.69
MINT	0	0	0	0	1	0	0	1	1	0	1	−1	−2.0	−0.76	−1.70	−0.69
PINT	0	0	0	0	0	1	0	1	1	0	1	+1	−1.0	−0.46	0.86	0.41
TINT	0	0	0	0	0	0	1	1	1	0	1	−1	−2.5	−0.85	−2.25	−0.81
FINE	0	1	0	0	0	0	0	1	1	1	0	+1	2.5	0.85	3.31	0.93
LINE	0	0	0	1	0	0	0	1	1	1	0	+1	2.0	0.76	2.52	0.85
MINE	0	0	0	0	1	0	0	1	1	1	0	+1	2.0	0.76	2.52	0.85
PINE	0	0	0	0	0	1	0	1	1	1	0	+1	3.0	0.91	5.09	0.98

Note: "Net" is the net input of the output unit; "Act" is its activation.

Table 6: Weights from Letter Units to Output Unit After Hebb Rule and Delta Rule Training

	Letter Units										
	D	F	H	L	M	P	T	I	N	E	T
Hebb Rule	−0.50	0.50	−0.50	0.00	0.00	1.00	−0.50	0.00	0.00	2.00	−2.00
Delta Rule	−0.84	0.59	−0.77	−0.19	−0.18	2.37	−0.73	0.24	0.24	2.23	−1.99

What has happened is that PINT's neighbors have cast slightly more votes for /i/ than for /I/.

We now consider results obtained using the delta rule. In this case, we trained the network for 20 epochs, again with a learning rate of 0.1. The overall magnitude of the weights is comparable to the Hebb rule case with only 5 epochs, since with the delta rule, the weight changes get smaller as the error gets smaller, and so the cumulative effect generally tends to be less More importantly, though, when the delta rule is used, the same general effects of consistency are observed, but now the response to PINT, though weaker than other responses, has the right sign. The reason for this is that the cumulative weight changes caused by PINT are actually larger than those caused by other items, because after the first epoch, the error is larger for PINT than for other items. Error-correcting learning eventually compensates for this but, before learning has completely converged, the effects of consistency are still apparent.

The error-correcting learning process causes an alteration in the relative weighting of the effects of neighbors, by assigning greater relative weight to those aspects of each input pattern that differentiate it from inconsistent patterns (see Table 6). This is why the weight tends to accumulate on P, which distinguishes PINT from the inconsistent neighbors DINT, HINT, LINT, MINT, and TINT. Correspondingly, the weights for D, H, and T are slightly more negative, to accentuate the differentiation of DINT, HINT, and TINT from PINT. The effect of consistency, then, is still present when the delta rule is used but, precisely because it makes the biggest changes where the errors are greatest, the delta rule tends to counteract the consistency effect.

Given that, with error-correcting learning, weight changes are proportional to the amount of error remaining, these changes decelerate as the network approaches zero error. This provides a second basis for the frequency-by-consistency interaction, over an above the effects of the sigmoid function. As Stone (1986) and Van Orden (1987) have pointed out, as error is reduced more on high- than on low-frequency items, there is less room for further relative improvement due to consistency among the former than among the latter.

For some tasks, including English word reading, no set of weights in a two-layer network will work for all of the training patterns (see Minsky & Papert, 1969). In such cases, hidden units that mediate between the input and output units are needed to achieve adequate performance.[7] Things are considerably more complex in networks with hidden

[7] An alternative strategy for increasing the range of tasks that can be solved by a two-layer network is to add additional input units that explicitly code relevant combinations of the original input units (see Gluck & Bower, 1988; Marr, 1969; Rumelhart, Hinton, & Williams, 1986a, for examples). In the domain of word reading, such higher-order units have been hand-specified by the experimenter as input units (Norris, submitted), hand-specified but derived from the input units as a separate pathway (Reggia, Marsland, & Berndt, 1988), or learned as hidden units in a separate

units, but Equation 13 still provides some guidance. The complexity comes from the fact that, for an output unit, $\mathcal{O}^{[pt]}$ reflects the similarities of the patterns of activation for training pattern p and test pattern t over the hidden units rather than over the input units. Even so, hidden units have the same tendency as output units to give similar output to similar inputs, as they use the same activation function. In fact, Equation 13 applies to them as well if $\delta_j^{[p]}$ is interpreted as the partial derivative of the error over all output units with respect to the summed input to the hidden unit j. The values of particular weights and the nonlinearity of the activation function can make hidden units relatively sensitive to some dimensions of similarity and relatively insensitive to others, and can even allow hidden units to respond to particular combinations of inputs and not to other, similar combinations. Such hidden units serve to reduce the similarity of the hidden representations of items, such as some orthographic neighbors, that differ in the presence/absence of these combinations. This is critical for learning complex mappings like the English spelling-to-sound system. Phoneme units respond on the basis of hidden-layer similarity, and they must respond quite differently to exception words than to their inconsistent neighbors in order for all of them to be pronounced correctly. Thus, by altering the similarities among input patterns, a network with hidden units can overcome the limitations of one with only input and output units. The process of learning to be sensitive to relevant input combinations occurs relatively slowly, however, because it goes against the network's inherent tendency toward making similar responses to similar inputs.

In summary, a broad range of connectionist networks exhibit the general trends that have been observed in human experimental data: robust consistency effects that tend to diminish with experience, both with specific items (i.e., frequency) and with the entire ensemble of patterns (i.e., practice). These factors are among the most important determinants of the time it takes for people to read words aloud.

Balancing Frequency and Consistency

The results of these analyses concur with the findings in empirical studies and in the SM89 and feedforward network simulations: there is an effect of consistency that diminishes with increasing frequency. Furthermore, details of the analytic results are also revealing. In particular, the extent to which the effect of consistency is eliminated in high frequency words depends on just how frequent they are relative to words of lower frequency. In fact, this effect may help to explain the discrepancy between the findings in the feedforward network and those in the SM89 network—namely, the existence of consistency effects among high-frequency words in the former but not in the latter (and not generally in empirical studies). At first glance, it would appear that the pattern observed in the feedforward network matches one in which the high-frequency words are of lower frequency relative to the low-frequency words (e.g., a frequency of 10 in Figure 8) than in the SM89 network (e.g., a frequency of 20). This is not literally true, however, because the same (logarithmically compressed) word frequencies were used in the two simulations.

A better interpretation is that, in the feedforward network, the effect of *consistency* is stronger than in the SM89 network and, relative to this, the effect of frequency appears weaker. As described earlier, the orthographic and phonological representations used by SM89, based on context-sensitive triples of letters and phonemes, disperse the regularities between the written and spoken forms of words. This has two relevant effects in the current context. The first is to reduce the extent to which the training on a given word improves performance on other words that share the same spelling-sound correspondences, and impairs performance on words that violate the correspondences. As illustrated earlier with the words LOG, GLAD, and SPLIT, even though a correspondence may be the same in a set of words, they may activate different orthographic and phonological units. As mentioned above, the weight changes induced by one word with help another only to the extent that they activate similar units (i.e., as a function of their overlap $\mathcal{O}^{[pt]}$). This effect is particularly important for low-frequency regular words, for which performance depends primarily on support from higher frequency words rather than from training on the word itself. In contrast, the new representations condense the regularities between orthography and phonology, so that weight changes for high-frequency words also improve performance on low-frequency with the same spelling-sound correspondences to a greater extent. Thus, there is an effect of frequency among regular words in the SM89 network but not in the feedforward network. For the same reason, in the SM89 network, performance on an exception word is less hindered by training on regular words that are inconsistent with it. It is almost as if regular words in the SM89 network behave like regular inconsistent words in the feedforward network, and exception words behave like ambiguous words: the support or interference they receive from similar words is somewhat reduced (see Figure 9).

The second way in which the SM89 representations serve to reduce the effect of consistency is to indirectly improve performance on exception words. This arises because the orthographic representations contain units that

pathway (Zorzi, Houghton, & Butterworth, 1994, in preparation).

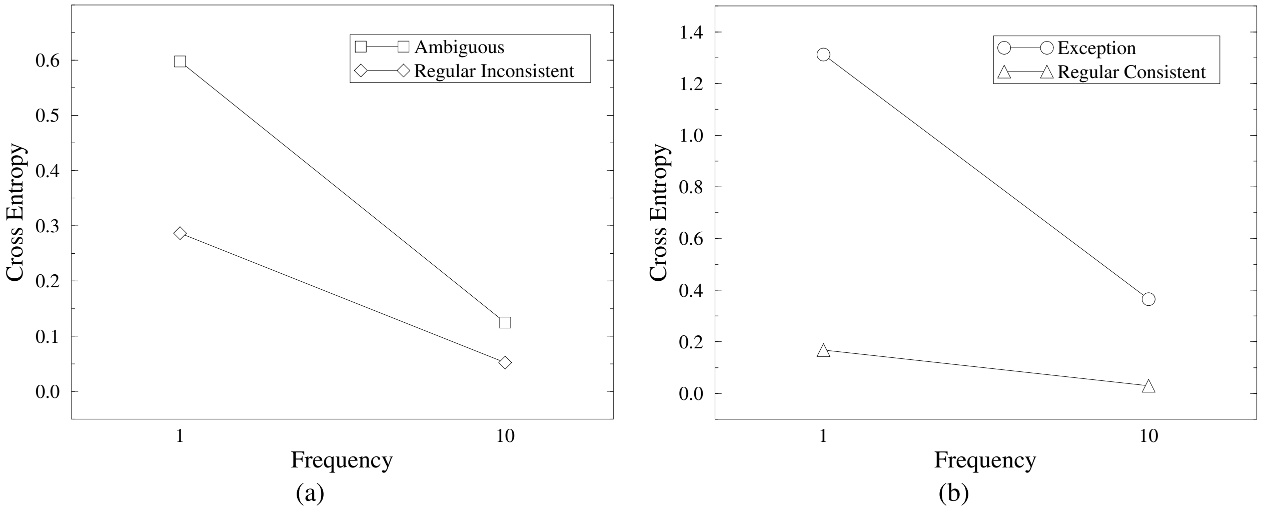

Figure 9: Data from the frequency-consistency equation (Equation 12 and Figure 8) for test words of frequencies 1 and 10, plotted separately for (a) regular inconsistent and ambiguous words and (b) regular consistent and exception words (right). The pattern in (a) is similar to that found for regular and exception words in the SM89 network (see Figure 2) while the one in (b) is similar to the pattern for the feedforward network (see Figure 5). The correspondences are only approximate due to the simplifying assumptions of the frequency-consistency equation.

explicitly indicate the presence of context-sensitive triples of letters. Some of these triples correspond to onset-vowel combinations and to word bodies (e.g., PIN, INT) that can directly contribute to the pronunciation of exception words (PINT). In contrast, although the new orthographic representations contain multiletter graphemes, none of them include both consonants and vowels, or consonants from both the onset and coda. Thus, for example, the orthographic units for P, I, N, and T contribute *independently* to the hidden representations. It is only at the hidden layer that the network can develop context-sensitive representations in order to pronounce exception words correctly, and it must learn to do this only on the basis of its exposure to words of varying frequency.

Nonetheless, it remains true that pattern of frequency and consistency effects in the SM89 network better replicates the findings in empirical studies than does the pattern in the feedforward network. Yet the same skilled readers exhibit a high level of proficiency at reading nonwords that is not matched in the SM89 network, but only in one using alternative representations that better capture the spelling-sound regularities. How can the effect of frequency and consistency be reconciled with good nonword reading?

The answer may lie in the fact that both the SM89 and the feedforward networks were trained using word frequency values that are logarithmically compressed from their true frequencies of occurrence in the language. Thus, the SM89 network replicates the empirical naming latency pattern because it achieves the appropriate *balance* between the influence of frequency and that of consistency, although both are suppressed relative to the effects in subjects. This suppression is revealed when nonword reading is examined, because on this task it is primarily the network's sensitivity to consistency that dictates performance. In contrast, by virtue of the new representations, the feedforward network exhibits a sensitivity to consistency that is comparably to that of subjects, as evidenced by its good nonword reading. But now, using logarithmic frequencies, the effects of frequency and consistency are unbalanced in the network and it fails to replicate the precise pattern of naming latencies of subjects.

This interpretation leads to the prediction that the feedforward network should exhibit both good nonword reading and the appropriate frequency and consistency effects if it is trained on words using their actual frequencies of occurrence. The next simulation tests this prediction.

Simulation 2: Feedforward Network with Actual Frequencies

The most frequent word in the Kucera and Francis (1967) list, THE, has a frequency of 69971 per million, while the least frequent words have a frequency of 1 per million. In the training procedure used by SM89, the probability that a word was presented to the network for training was proportional to the logarithm of its frequency rather than its actual frequency. This compresses the effective frequency range from about 70000:1 to about 16:1. Thus, the network experiences much less variation in the frequency of occurrence of words than do normal readers.

SM89 put forward a number of arguments in favor of using logarithmically compressed frequencies rather than actual frequencies in training their network. Beginning readers have yet to experience enough words to approximate the actual frequency range in the language. Also, low-frequency words disproportionately suffer from the lack of inflectional and derivational forms in the training corpus. However, the main reason for compressing the frequency range was a practical consideration based on limitations of the available computational resources. If the highest frequency word was presented every epoch, the lowest frequency words would be presented on average only about every 70,000 epochs. Thus, if actual frequencies were used, SM89 could not have trained their network long enough for it to have had sufficient exposure on low-frequency words.

To compound matters, as SM89 point out, basic properties of the network and training procedure already serve to progressively weaken the impact of frequency over the course of training. In an error-correcting training procedure like back-propagation, weights are changed only to the extent that doing so reduces the mismatch between the generated and correct output. As high-frequency words become mastered, they produce less mismatch and so induce progressively smaller weight changes. This effect is magnified by the fact that, due to the asymptotic nature of the unit input-output function, weight changes have smaller and smaller impact as units approach their correct extremal values. As a result, learning becomes dominated mostly by lower frequency words that are still inaccurate, effectively compressing the range of frequency driving learning in the network.

Thus, SM89 considered it important to verify that their results did not depend critically on the use of such a severe frequency compression. They trained a version of the network in which the probability that a word is presented during an epoch is based on the square-root of its frequency rather than the logarithm (resulting in a frequency range of about 265:1 rather than 16:1). They found the same basic pattern of frequency and consistency effects in naming latency for the Taraban and McClelland (1987) words, although there was a larger effect of frequency among regular words, and virtually no effect of consistency among high-frequency words even early in training. This shift corresponds predictably to a pattern in which the influence of frequency is stronger relative to the influence of consistency. However, SM89 present no data on the network's accuracy in reading words or nonwords.

In the current simulation, we train a version of the feedforward network (with the new representations) using the actual frequencies of occurrence of words. The training procedure in the current work avoids the problem of sampling low-frequency words by using frequency to directly scale the weight changes induced by a word—this is equivalent to sampling in the limit of a small learning rate, and it allows any range of frequencies to be employed. The goal is to test the hypothesis that, by balancing the strong influence of consistency that arises from the use of representations that better capture spelling-sound regularities with a realistically strong influence of frequency, the network should exhibit the appropriate pattern of frequency and consistency effects in naming latency while also producing accurate performance on word and nonword pronunciation.

Method

Network Architecture

The architecture of the network is the same as in the Simulation 1 (see Figure 3).

Training Procedure

The only major change in the training procedure from Simulation 1 is that, as described above, the values used to scale the error derivatives computed by back-propagation are proportional to the actual frequencies of occurrence of the words (Kucera & Francis, 1967) rather than to a logarithmic compression of their frequencies. Following SM89, the 82 words in the training corpus that are not listed in Kucera and Francis (1967) were assigned a frequency of 2, and all others were assigned their listed frequency plus 2. These values were then divided by the highest value in the corpus (69973 for THE) to generate the scaling values used during training. Thus, the weight changes produced by the word THE are unscaled (i.e., scaling value of 1.0). For comparison, AND, the word with the next highest frequency (28860 occurrences per million), has a value of 0.412. By contrast, the relative frequencies of most other words is extremely low. The mean scaling value across the entire training corpus is 0.0020, while the median value is 0.00015. Taraban and McClelland's (1987) high-frequency exception words have an average value of 0.014 while the low-frequency exception words average 0.00036. Words not in the Kucera and Francis (1967) list have a value just under 3×10^{-5}.

In addition, two parameters of the training procedure were modified to compensate for the changes in word frequencies. First, the global learning rate, ϵ in Equation 5, was increased from 0.001 to 0.05, to compensate for the fact that the summed frequency for the entire training corpus is reduced from 683.4 to 6.05 when using actual rather than logarithmic frequencies. Second, the slight tendency for weights to decay towards zero was removed, to prevent the very small weight changes induced by low-frequency words (due to their very small scaling factors) from being overcome by the tendency of weights to shrink towards zero.

Other than for these modifications, the network was trained in exactly the same way as in Simulation 1.

Testing Procedure

The procedure for testing the network's procedure on words and nonwords is the same as in Simulation 1.

Results

Word Reading

As the weight changes caused by low-frequency words are so small, considerably more training is required to reach approximately the same level of performance as when using logarithmically compressed frequencies. After 1300 epochs of training, the network mispronounces only 7 words in the corpus: BAS, BEAU, CACHE, CYST, GENT, TSAR, and YEAH (99.8% correct).[8] These words have rather inconsistent spelling-sound correspondences and have very low frequencies (i.e., an average scaling value of 9.0×10^{-5}). Thus, the network has mastered all of the exception words except a few of the very lowest in frequency.

Nonword Reading

Table 7 lists the errors made by the network in pronouncing the lists of nonwords from Glushko (1979) and from McCann and Besner (1987). The network produces "regular" responses to 42/43 (97.7%) of Glushko's regular nonwords, 39/43 (67.4%) of the exception nonwords, and 66/80 (82.5%) of McCann and Besner's control nonwords. Using a criterion that more closely corresponds to that used with subjects—considering a response correct if it is consistent with the pronunciation of a word in the training corpus (and not considering inflected words or those with J in the coda)—the network achieves 42/43 (97.7%) correct on both the regular and exception nonwords, and 68/76 (89.5%) correct on the control nonwords. Thus, the network's performance on these sets of nonwords is comparable to that of subjects and to that of the network trained on logarithmic frequencies.

Frequency and Consistency Effects

Figure 10 shows the mean cross entropy error of the network in pronouncing words of varying degrees of spelling-sound consistency (Taraban & McClelland, 1987) as a function of frequency. There is a main effect of frequency ($F_{1,184}$=22.1, $p<.001$), a main effect of consistency ($F_{3,184}$=6.49, $p<.001$), and an interaction of frequency and consistency ($F_{1,184}$=5.99, $p<.001$). Post hoc comparisons show that the effect of frequency is significant at the 0.05 level among words of each level of consistency when considered separately.

The effect of consistency is significant among low frequency words ($F_{3,92}$=6.25, p=.001) but not among high-frequency words ($F_{3,92}$=2.48, p=.066). Post hoc comparisons among low-frequency words revealed that the difference in error between exception words and ambiguous words is significant ($F_{1,46}$=4.09, p=.049), the difference between regular consistent and inconsistent words is marginally significant ($F_{1,46}$=3.73, p=.060), but the difference between ambiguous words and regular inconsistent words fails to reach significance ($F_{1,46}$=2.31, p=.135).

Overall, this pattern of results matches the one found in empirical studies fairly well. Thus, by balancing the influence of frequency and consistency in the network, it replicates the pattern of interaction of these variables on naming latency while also reading words and nonwords as accurately as skilled readers.

[8] Homographs are considered correct if they elicit either correct pronunciation.

Table 7: Errors by the feedforward network trained with actual frequencies in pronouncing nonwords from Glushko (1979) and McCann and Besner (1987).

Glushko (1979)			McCann and Besner (1987)		
Nonword	Correct	Response	Nonword	Correct	Response
Regular Nonwords (1/43)			Control Nonwords (14/80)		
*WOSH	/waS/	/woS/	TUNCE	/t∧ns/	/tUns/
Exception Nonwords (14/43)			*TOLPH	/tolf/	/tOl(f 0.13)/
BLEAD	/blEd/	/bled/	*ZUPE	/zUp/	/(z 0.09)yUp/
BOST	/bost/	/bOst/	SNOCKS	/snaks/	/snask(ks 0.31)/
COSE	/kOz/	/kOs/	*GOPH	/gaf/	/gaT/
GROOK	/grUk/	/gruk/	*VIRCK	/vurk/	/(v 0.13)urk/
*HEAF	/hEf/	/h@f/	LOKES	/lOks/	/lOsk(ks 0.00)/
HOVE	/hOv/	/h∧v/	*YOWND	/yWnd/	/(y 0.04)and/
LOME	/lOm/	/l∧m/	KOWT	/kWt/	/kOt/
PILD	/pild/	/pIld/	*FUES	/fyUz/	/fyU(z 0.45)/
PLOVE	/plOv/	/pl∧v/	*HANE	/hAn/	/h@n/
POOT	/pUt/	/put/	FAIJE	/fAj/	/fA(j 0.00)/
POVE	/pOv/	/p∧v/	*ZUTE	/zUt/	/(z 0.01)yUt/
SOOD	/sUd/	/sud/	JINJE	/jinj/	/jIn(j 0.00)/
WEAD	/wEd/	/wed/			
WONE	/wOn/	/w∧n/			

Note: The activity levels of correct but missing phonemes are listed in parentheses. In these cases, the actual response is what falls outside the parentheses. Words marked with "*" remain errors after considering properties of the training corpus (as explained in the text).

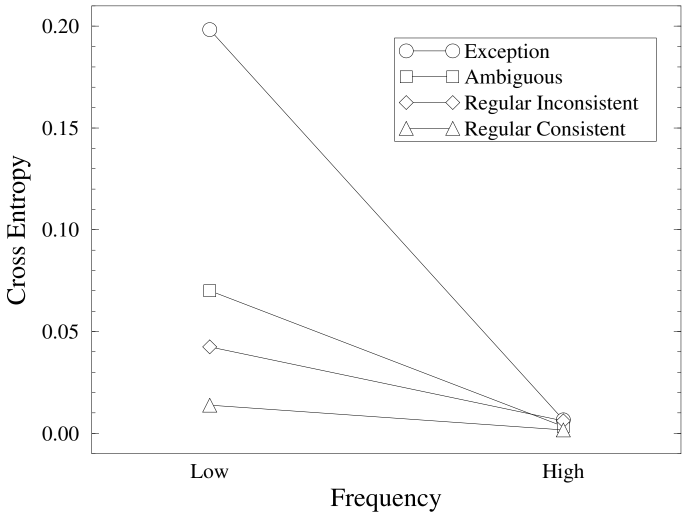

Figure 10: Mean cross-entropy error produced by the feedforward network trained on actual frequencies for words with various degrees of spelling-sound consistency as a function of frequency.

Training with a Moderate Frequency Compression

As SM89 argued, training with the actual frequencies of monosyllabic words might not provide the best approximation to the experience of readers. In particular, most multisyllabic words have consistent spelling-sound correspondences, so that training only on monosyllabic words underestimates a reader's exposure to spelling-sound regularities. Training with a compressed frequency range compensates for this bias because exception words tend to be of higher frequency than regular words and, thus, are disproportionately affected by the compression.

We have seen that a very severe (logarithmic) compression reduces the effect of frequency to such an extent that a network using representations that amplify consistency effects fails to exhibit the exact pattern of naming latencies found in empirical studies. Nonetheless, it would seem appropriate to test whether a less severe compression results in a better match to the empirical findings. As mentioned earlier, SM89 found that presenting words during training with a probability proportional to the square-root of their frequency replicates the basic frequency and consistency effects in their network, but they presented no data on the accuracy of the network's performance. Accordingly, it seemed worthwhile for comparison purposes to train a network with the new representations also using a square-root compression of word frequencies.

Analogous to the use of actual frequencies, the scaling value for each word was the square-root of its Kucera and Francis (1967) frequency plus 2, divided by the square-root of the frequency of THE plus 2 (264.5). The value for AND is 0.642. The mean for the corpus is 0.023 and the median is 0.012. Taraban and McClelland's (1987) high-frequency exception words average 0.097 while the low-frequency exception words average 0.017. Words not in the Kucera and Francis (1967) list have a value of 0.0053. Thus, the compression of frequency is much less severe than when using logarithms but it is still substantial.

The summed frequency of the training corpus is 69.8; accordingly, the global learning rate, ϵ, was adjusted to 0.01. The training procedure is otherwise identical to that used when training on the actual word frequencies.

Word Reading

After 400 epoch, the network pronounces correctly all words in the training corpus except HOUSE, for which the states of both the final /s/ and the final /z/ just fail to be active (/s/: 0.48, /z/: 0.47). Thus, the network's word reading is essentially perfect.

Nonword Reading

The network makes no errors on Glushko's (1979) regular nonwords. On the exception nonwords, 14 of the network's responses are nonregular, but all but one of these (POVE ⇒/pav/) are consistent with some word in the training corpus (97.7% correct). The network mispronounces 13 of McCann and Besner's (1987) control nonwords. However, only 7 of these remain as errors when using the same scoring criterion as was used with subject and ignoring inflected forms and those with final J (90.8% correct). Thus, the network trained with square-root frequencies pronounces nonwords as well, if not slightly better, than the network trained with actual frequencies.

Frequency and Consistency Effects

Figure 11 shows the mean cross entropy error of the network in pronouncing words of varying degrees of spelling-sound consistency (Taraban & McClelland, 1987) as a function of frequency. Overall, there is a significant effect of frequency ($F_{1,184}$=47.7, p<.001), consistency, ($F_{1,184}$=14.9, p<.001), and interaction of frequency and consistency ($F_{3,184}$=8.409, p<.001). The effect of frequency is also significant at the 0.05 level among words of each level of consistency when considered separately. Among high-frequency words, regular inconsistent, ambiguous, and exception words are significantly different from regular consistent words but not from each other. Among low-frequency words, the difference between regular inconsistent words and ambiguous words is not significant ($F_{1,46}$=1.18, p=.283) but all other pairwise comparisons are. Thus, this network also replicates the basic empirical findings of the effects of frequency and consistency on naming latency.

Summary

The SM89 simulation replicates the empirical pattern of frequency and consistency effects by appropriately balancing the relative influences of these two factors. Unfortunately, both are reduced relative to their strength in skilled

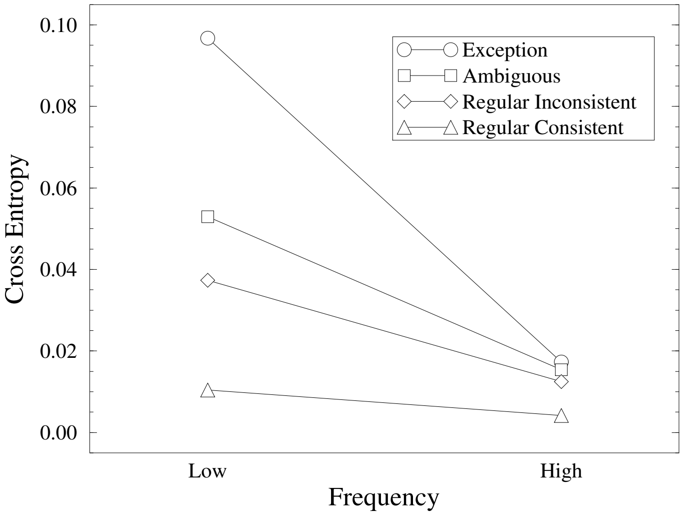

Figure 11: Mean cross-entropy error produced by the feedforward network trained on square-root frequencies for words with various degrees of spelling-sound consistency as a function of frequency.

readers. The fact that the orthographic and phonological representations disperse the regularities between spelling and sound serves to diminish the relative impact of consistency. Likewise, the use of a logarithmic compression of the probability of word presentations serves to diminish the impact of frequency. As a result of the reduced effectiveness of consistency, nonword reading suffers.

The current work uses representations that better capture spelling-sound regularities, thereby increasing the relative influence of consistency. One effect of this is to improve nonword reading to a level comparable to that of skilled readers. However, if a logarithmic frequency compression continues to be used, the relative impact of frequency is too weak and the network exhibits consistency effects among high-frequency words not found in empirical studies.

The appropriate relative balance of frequency and consistency can be restored, while maintaining good nonword reading, by using the actual frequencies of words during training. In fact, a square-root frequency compression that is much more moderate that a logarithmic one also replicates the empirical naming latency pattern, although a consistency effect among high-frequency words begins to emerge. In this way, the three network presented thus far—trained on logarithmic frequencies, square-root-frequencies, or actual frequencies—provide clear points of comparison of the relative influences of word frequency and spelling-sound consistency on naming latency. Together with the analytical results from the previous section, the findings suggest that these basic findings in word and nonword reading can be interpreted naturally in terms of the basic principles of operation of connectionist networks that are exposed to an appropriately structured training corpus.

Simulation 3: Interactivity, Componential Attractors, and Generalization

As outlined earlier, the current approach to lexical processing is based on a number of general principles of information processing, loosely expressed by the acronym GRAIN (for Graded, Random, Adaptive, Interactive, and Nonlinear). Together with the principles of distributed representations and knowledge, the approach constitutes a substantial departure from traditional assumptions about the nature of language knowledge and processing (e.g., Pinker, 1991). It must be noted, however, that the simulations presented so far involve only deterministic, feedforward networks, and thus fail to incorporate two important principles: interactivity and randomness (intrinsic variability). In part, this simplification has been necessary for practical reasons; interactive, stochastic simulations are far more demanding of computational resources. More importantly, including only some of the relevant principles in a given simulation enables the specific contribution that each makes to the overall behavior of the system to be analyzed in greater detail. This has been illustrated most clearly in the current work with regard to the nature of the distributed representations used for orthography and phonology, and the relative influences of frequency and consistency on network learning (adaptivity). Nonetheless, each such network constitutes only an approximation or abstraction of a more complete simulation that would incorporate all of the principles. The methodology of considering sets of principles separately implicitly assumes that there are no unforeseen, problematic interactions among the principles, such that the findings with simplified simulations would not generalize to more comprehensive ones.

The current simulation investigates the implications of interactivity for the process of pronouncing written words and nonwords. Interactivity plays an important role in connectionist explanations of a number of cognitive phenomena (McClelland & Elman, 1986; McClelland & Rumelhart, 1981; McClelland, 1987), and constitutes a major point of contention with alternative theoretical formulations (Massaro, 1988, 1989). In a network, processing is interactive when units can mutually constrain each other in settling on the most consistent interpretation of the input. For this to be possible, the architecture of the network must be generalized to allow feedback or *recurrent* connections among units. For example, in the interactive activation model of letter and word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), letter units and word units are bidirectionally connected so that the partial activation of a word unit can feed back to support the activation of letter units with which it is consistent.

A common way in which interactivity has been employed in networks is in making particular patterns of activity into stable *attractors*. In an attractor network, units interact and update their states repeatedly in such a way that the initial pattern of activity generated by an input gradually settles to the nearest attractor pattern. A useful way of conceptualizing this process is in terms of a multidimensional *state* space in which the activity of each unit is plotted along a separate dimension. At any instant in time, the pattern of activity over all of the units corresponds to a single point in this space. As units change their states in response to a given input, this point moves in state space, eventually arriving at the (attractor) point corresponding to the network's interpretation. The set of initial patterns that settle to this same final pattern corresponds to a region around the attractor, called its *basin* of attraction. To solve a task, the network must learn connection weights that cause units to interact in such a way that the appropriate interpretation of each input is an attractor whose basin contains the initial pattern of activity for that input.

In the domain of word reading, attractors have played a critical role in connectionist accounts of the nature of normal and impaired reading via meaning (Hinton & Sejnowski, 1986; Hinton & Shallice, 1991; Plaut & Shallice, 1993). According to these accounts, the meanings of words are represented in terms of patterns of activity over a large number of semantic features.[9] As only a small fraction of the possible combinations of features correspond to the meanings of actual words, it is natural for a network to learn to make these semantic patterns into attractors. Then, in deriving the meaning of a word from its orthography, the network need only generate a initial pattern of activity that falls somewhere within the appropriate semantic attractor basin; the settling process will clean up this pattern into the exact meaning of the word.[10] If, however, the system is damaged, the initial activity for a word may fall within a neighboring attractor basin, typically corresponding to a semantically-related word. The damaged network will then settle to the exact meaning of that word, resulting in a semantic error (e.g., CAT read as "dog"). In fact, the occurrence of such errors is the hallmark symptom of a type of acquired reading disorder known as *deep dyslexia* (see Coltheart, Patterson, & Marshall, 1980, for more details on the full range of symptoms of deep dyslexia, and Plaut & Shallice, 1993, for connectionist simulations replicating these symptoms). In this way, attractors obviate the need for word-specific units in mediating between orthography and semantics (see Hinton, McClelland, & Rumelhart, 1986, for discussion).

When applied to the mapping from orthography to phonology, however, the use of interactivity to form attractors would appear problematic. In particular, the correct pronunciation of a nonword typically does not correspond to the pronunciation of some word. If the network develops attractors for word pronunciations, one might expect that the input for a nonword would often be captured within the attractor basin for a similar word, resulting in many incorrect *lexicalizations*. More generally, attractors would seem to be appropriate only for tasks, such as semantic categorization or object recognition, in which the correct response to a novel input is a familiar output. By contrast, in oral reading, the correct response to a novel input is often a novel output. If it is true that attractors cannot support this latter sort of generalization, their applicability in reading specifically, and cognitive science more generally, would be fundamentally limited.

The current simulation demonstrates that these concerns are ill-founded, and that, with appropriately structured representations, the principle of interactivity can operate effectively in the phonological pathway as well as in the semantic pathway (see Figure 1). The reason is that, in learning to map orthography to phonology, the network develops

[9] More complex, frame-like representations can be implemented using this approach if units can represent conjunctions of roles and properties of role-fillers (Hinton, 1981; Derthick, 1990).

[10] This characterization of deriving word meanings is necessarily oversimplified. Words with multiple, distinct meanings would map to one of a number of separate semantic attractors. Shades of meaning across contexts could be expressed by semantic attractors that are *regions* in semantic space instead of single points. Notice that these two conditions can be seen as ends of a continuum involving various degrees of similarity and variability among the semantic patterns generated by a word across contexts (also see McClelland, St. John, & Taraban, 1989).

attractors for words that are *componential*—they have substructure that reflects common sublexical correspondences between orthography and phonology. This substructure applies not only to words but also to nonwords, enabling them to be pronounced correctly. At the same time, the network develops less componential attractors for words with exceptional spelling-sound correspondences. Thus, rather than being a hindrance, attractors are a particularly effective style of computation for quasi-regular tasks such as word reading.

A further advantage of an attractor network over a feedforward network in modeling word reading is that the former provides a more direct analogue of naming latency. Thus far, we have followed SM89 in using an error measure in a feedforward network to account for naming latency data from subjects. SM89 offer two justifications for this approach. The first is based on the assumption that the accuracy of the phonological representation of a word would directly influence the execution speed of the corresponding articulatory motor program (see Lacouture, 1989; Zorzi et al., 1994, in preparation, for simulations embodying this assumption). This assumption is consistent with the view that the time required by the orthography-to-phonology computation itself does not vary systematically with word frequency or spelling-sound consistency. If this were the case, a feedforward network of the sort SM89 and we have used, which takes the same amount of time to process any input, would be a reasonable rendition of the nature of the phonological pathway in subjects.

An alternative justification for the use of error scores to model naming latencies, mentioned only briefly by SM89, is based on the view that the actual computation from orthography to phonology involves interactive processing, such that the time to settle on an appropriate phonological representation does vary systematically with word type. The naming latencies exhibited by subjects are a function of this settling time, perhaps combined with articulatory effects. Accordingly, a feedforward implementation of the mapping from orthography to phonology should be viewed as an simplification of a recurrent implementation that would more accurately approximate the actual word reading system. Studying the feedforward implementation is still informative because many of its properties, including its sensitivity to frequency and consistency, depend on computational principles of operation that would also apply to a recurrent implementation—namely, adaptivity, distributed representations and knowledge, and nonlinearity. These principles merely manifest themselves differently: influences that reduce error in a feedforward network serve to accelerate settling in a recurrent network. Thus, error in a feedforward network is a valid approximation of settling time in a recurrent network because they both arise from the same underlying causes: additive frequency and consistency effects in the context of a nonlinear gradual ceiling effect. Nonetheless, even given these arguments, it seems worthwhile to verify that settling time in a recurrent implementation can model the relevant empirical pattern of naming latencies as well as error in a feedforward implementation.

Method

Network Architecture

The architecture of the attractor network is shown in Figure 12. The numbers of grapheme, hidden, and phoneme units are the same as in the feedforward networks, but the attractor network has some additional sets of connections. Each input unit is still connected to each hidden unit which, in turn, is connected to each phoneme unit. In addition, each phoneme unit is connected to each other phoneme unit (including itself), and each phoneme unit sends a connection back to each hidden unit. The weights on the two connections between a pair of units (e.g., a hidden unit and a phoneme unit) are trained separately and need not have identical values. Including the biases of the hidden and phoneme units, the network has a total of 26,582 connections.

The states of units in the network change smoothly over time in response to influences from other units. In particular, the instantaneous change over time t of the input x_j to unit j is proportional to the difference between its current input and the summed contribution from other units.

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \left(\sum_i s_i w_{ij} \; - \; x_j \right) \tag{14}$$

The state s_j of unit j is $\sigma(x_j)$, the standard logistic function of its integrated input, that ranges between 0.0 and 1.0 (see Equation 2). For clarity, we will call the summed input from other units i the *external* input to each unit, to distinguish it from the *integrated* input that governs the unit's state.

According to Equation 14, when a unit's integrated input is perfectly consistent with its external input (i.e., $x_j = \sum_i s_i w_{ij}$), the derivative is zero and the unit's integrated input, and hence its state, ceases to change. Notice that its

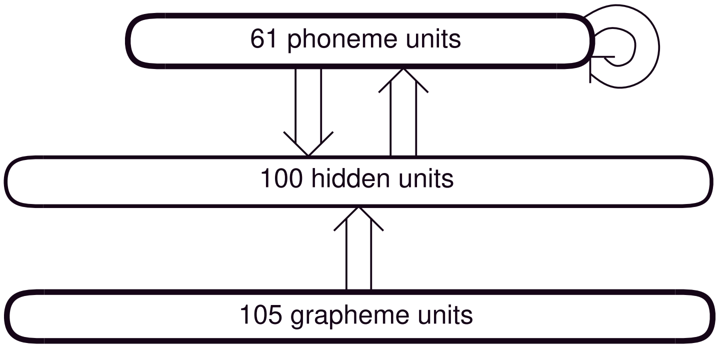

Figure 12: The architecture of the attractor network. Ovals represent groups of units, and arrows represent complete connectivity from one group to another.

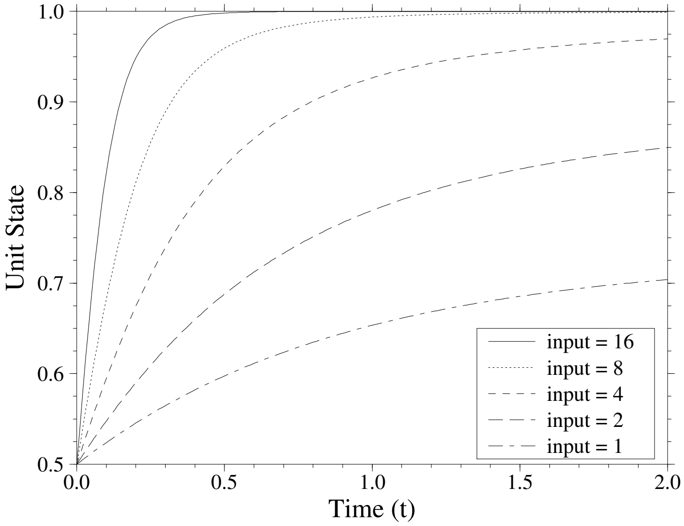

Figure 13: The state over time of a continuous unit, initialized to 0.5 and governed by Equation 14, when presented with fixed external input from other units of varying magnitude. The curves of state values for negative external input are the exact mirror images of these curves, approaching 0.0 instead of 1.0.

activity at this point, $\sigma\left(\sum_i s_i w_{ij}\right)$, is exactly the same as it would be if it were a standard unit that computes its state from the external input instantaneously (as in a feedforward network; see Equations 1 and 2). To illustrate this, and to provide some sense of the temporal dynamics of units in the network, Figure 13 shows the activity over time of a single unit, initialized to 0.5 and governed by Equation 14, in response to external input of varying magnitude. Notice that, over time, the unit state gradually approaches an asymptotic value equal to the logistic function applied its external input.

For the purposes of simulation on a digital computer, it is convenient to approximate continuous units with finite difference equations, in which time is discretized into *ticks* of some duration τ:

$$\triangle x_j = \tau \left(\sum_i s_i w_{ij} \; - \; x_j \right)$$

where $\triangle x_j = x_j^{[t]} - x_j^{[t-\tau]}$. Using explicit superscripts for discrete time, this can be rewritten as

$$x_j^{[t]} = \tau \sum_i s_i w_{ij} + (1 - \tau)\, x_j^{[t-\tau]} \qquad (15)$$

According to this equation, a unit's input at each time tick is a weighted average of its current input and that dictated by other units, where τ is the weighting proportion.[11] Notice that, in the limit (as $\tau \rightarrow 0$) this discrete computation becomes identical to the continuous one. Thus, adjustments to τ affect the accuracy with which the discrete system approximates the continuous one, but do not alter the underlying computation being performed. This is of considerable practical importance, as the computational time required to simulate the system is inversely proportional to τ. A relatively larger τ can be used during the extensive training period (0.2 in the current simulation), when minimizing computation time is critical, whereas a much smaller τ can be used during testing (e.g., 0.01), when a very accurate approximation is desired. As long as τ remains sufficiently small for the approximations to be adequate, these manipulations do not fundamentally alter the behavior of the system.

Training Procedure

The training corpus for the network is the same as used with the feedforward network trained on actual word frequencies. As in that simulation, the frequency value of each word is used to scale with weight changes induced by the word.

The network is trained with a version of back-propagation designed for recurrent networks, known as *back-propagation through time* (Rumelhart et al., 1986a; Williams & Peng, 1990), further adapted for continuous units (Pearlmutter, 1989). In understanding back-propagation through time, it may help to think of the computation in standard back-propagation in a three layer feedforward network as occurring over time. In the forward pass, the states of input units are clamped at time $t = 0$. Hidden unit states are computed at $t = 1$ from these input unit states, and then output unit states are computed at $t = 2$ from the hidden unit states. In the backward pass, error is calculated for the output units based on their states ($t = 2$). Error for the hidden units and weight changes for the hidden-to-output connections are calculated based on the error of the output units ($t = 2$) and the states of hidden units ($t = 1$). Finally, the weight changes for the input-to-hidden connections are calculated based on the hidden unit error ($t = 1$) and the input unit states ($t = 0$). Thus, feedforward back-propagation can be interpreted as involving a pass forward in time to compute unit states, followed by a pass backward in time to compute unit error and weight changes.

Back-propagation through time has exactly the same form, except that, because a recurrent network can have arbitrary connectivity, each unit can receive contributions from any unit at any time, not just from those in earlier layers (for the forward pass) or later layers (for the backward pass). This means that each unit must store its state and error at each time tick, so that these values are available to other units when needed. In addition, the states of non-input units affect those of other units immediately, so they need to be initialized to some neutral value (0.5 in the current simulation). In all other respects, back-propagation through time is computationally equivalent to feedforward back-propagation. In fact, back-propagation through time can be interpreted as "unfolding" a recurrent network into a much larger feedforward network with a layer for each time tick composed of a separate copy of all the units in the recurrent network (see Minsky & Papert, 1969; Rumelhart et al., 1986a)

In order to apply back-propagation through time to continuous units, the propagation of error in the backward pass must be made continuous as well (Pearlmutter, 1989). If we use δ_j to designate the derivative of the error with respect to the input of unit j, and $\sigma'(\cdot)$ is the derivative of the logistic function, then, in feedforward back-propagation:

$$\delta_j = \frac{\partial E}{\partial s_j} \sigma'(x_j)$$

In the discrete approximation to back-propagation through time with continuous units, this becomes

$$\delta_j^{[t]} = \tau \frac{\partial E}{\partial s_j^{[t+\tau]}} \sigma'\left(x_j^{[t+\tau]}\right) \; + \; (1-\tau)\, \delta_j^{[t+\tau]}$$

Thus, δ_j is a weighted average backwards in time of its current value and the contribution from the current error of the unit. In this way, as in standard back-propagation, δ_j in the backward pass is analogous to x_j in the forward pass (cf. Equation 15).

As output units can interact with other units over the course of processing a stimulus, they can indirectly affect the

[11] These temporal dynamics are somewhat different from those of the Plaut and McClelland (1993, Seidenberg et al., in press) network. In that network, each unit's input was set instantaneously to the summed external input from other units; the unit's state was a weighted average of its current state and the one dictated by its instantaneous input.

error for other output units. As a result, the error for an output unit becomes the sum of two terms: the error due to the discrepancy between its own state and its target, and the error back-propagated to it from other units. The first term is often referred to as error that is *injected* into the network by the training environment, while the second term might be thought of as error that is *internal* to the network.

Given that the states of output units vary over time, they can have targets that specify what states they should be in at particular points in time. Thus, in back-propagation though time, error can be injected at any or all time ticks, not just at the last one as in feedforward back-propagation. Targets that vary over time define a trajectory that the output states will attempt to follow (see Pearlmutter, 1989, for a demonstration of this type of learning). If the targets remain constant over time, however, the output units will attempt to reach their targets as quickly as possible and remain there. In the current simulation, we use this technique to train the network to form stable attractors for the pronunciations of words in the training corpus.

It is possible for the states of units to change quickly if they receive a very large summed input from other units (see Figure 13). However, even for rather large summed input, units typically require some amount of time to approach an extremal value, and may never completely reach it. As a result, it is practically impossible for units to achieve targets of 0.0 or 1.0 immediately after a stimulus has been presented. For this reason, in the current simulation, a less stringent training regime is adopted. Although the network is run for 2.0 units of time, error is injected only for the second unit of time; units receive no direct pressure to be correct for the first unit of time (although back-propagated internal error causes weight changes that encourage units to move towards the appropriate states as early as possible). In addition, output units are trained to targets of 0.1 and 0.9 rather than 0.0 and 1.0, and no error is injected if a unit exceeds its target (e.g., reaches a state of 0.95 for a target of 0.9). This training regime can be achieved by units with only moderately large summed input (see the curve for input = 4 in Figure 13).

As with the feedforward network using actual frequencies, the attractor network was trained with a global learning rate $\epsilon = 0.05$ (with adaptive connection-specific rates) and momentum $\alpha = 0.9$. Furthermore, as mentioned above, the network was trained using a discretization $\tau = 0.2$. Thus, units update their states 10 times $(2.0/0.2)$ in the forward pass, and they back-propagate error 10 times in the backward pass. As a result, the computational demands of the simulation are about 10 times that of one of the feedforward simulations. In an attempt to reduce the training time, momentum was increased to 0.98 after 200 epochs. To improve the accuracy of the network's approximation to a continuous system near the end of training, τ was reduced from 0.2 to 0.05 at epoch 1800, and reduced further to 0.01 at epoch 1850 for an additional 50 epochs of training. During this final stage of training, each unit updated its state 200 times over the course of processing each input.

Testing Procedure

A fully adequate characterization of response generation in distributed connectionist networks would involve stochastic processing (see McClelland, 1991) and, thus, is beyond the scope of the present work. As an approximation in a deterministic attractor network, we use a measure of the time it takes the network to compute a stable output in response to a given input. Specifically, the network responds when the average change in the states of the phoneme units falls below some criterion (0.00005 with $\tau = 0.01$ for the results below).[12] At this point, the network's naming latency is the amount of continuous time that has passed in processing the input, and its naming response is generated on the basis of the current phoneme states using the same procedure as for the feedforward networks.

Results

Word Reading

After 1900 epochs of training, the network pronounces correctly all but 25 of the 2998 words in the training corpus (99.2% correct). About half of these are regularizations of low-frequency exception words (e.g., SIEVE ⇒/sEv/, SUEDE ⇒/swEd/, TOW ⇒/tW/). Most of the remaining errors would be classified as visual errors (e.g., FALL ⇒/folt/, SHRIEK ⇒/SrIk/, HASP ⇒/h@ps/) although a number of pronunciations are simply missing consonants (e.g., ACHE ⇒ /A/, BEIGE ⇒/bA/, TZAR ⇒/ar/). All in all, the network has come close to mastering the training corpus, although its performance is slightly worse than that of the equivalent feedforward network.

[12] This specific criterion was chosen because it gives rise to mean response times that are within the 2.0 units of time over which the network was trained; other criteria produce qualitatively equivalent results.

Table 8: Errors by the attractor network trained with actual word frequencies in pronouncing nonwords from Glushko (1979) and McCann and Besner (1987).

Glushko (1979)			McCann and Besner (1987)		
Nonword	Correct	Response	Nonord	Correct	Response
Regular Nonwords (3/43)			Control Nonwords (11/80)		
*HODE	/hOd/	/hOdz/	*KAIZE	/kAz/	/skwAz/
*SWEAL	/swEl/	/swel/	ZUPE	/zUp/	/zyUp/
*WOSH	/waS/	/wuS/	*JAUL	/jol/	/jOl/
Exception Nonwords (16/43)			*VOLE	/vOl/	/vOln/
BLEAD	/blEd/	/bled/	*YOWND	/yWnd/	/(y 0.04)Ond/
BOST	/bost/	/bOst/	KOWT	/kWt/	/kOt/
COSE	/kOz/	/kOs/	*VAWX	/voks/	/voNks/
COTH	/koT/	/kOT/	FAIJE	/fAj/	/fA(j 0.00)/
GROOK	/grUk/	/gruk/	ZUTE	/zUt/	/zyUt/
LOME	/lOm/	/l∧m/	*YOME	/yOm/	/yam/
MONE	/mone/	/m∧n/	JINJE	/jinj/	/jIn(j 0.00)/
PLOVE	/plOv/	/plUv/			
POOT	/pUt/	/put/			
*POVE	/pOv/	/pav/			
SOOD	/sUd/	/sud/			
SOST	/sost/	/sOst/			
SULL	/s∧l/	/sul/			
WEAD	/wEd/	/wed/			
WONE	/wOn/	/w∧n/			
WUSH	/w∧S/	/wuS/			

Note: The activity levels of correct but missing phonemes are listed in parentheses. In these cases, the actual response is what falls outside the parentheses. Words marked with "*" remain errors after considering properties of the training corpus (as explained in the text).

Nonword Reading

Table 8 lists the errors made by the network in pronouncing the lists of nonwords from Glushko (1979) and from McCann and Besner (1987). The network produces "regular" pronunciations to 40/43 (93.0%) of Glushko's regular nonwords, 27/43 (62.8%) of the exception nonwords, and 69/80 (86.3%) of McCann and Besner's control nonwords. If we accept as correct any pronunciation that is consistent with that of a word in the training corpus with the same body (and ignore inflected words and those with final J), the network pronounces correctly 42/43 (97.7%) of the exception nonwords, and 70/76 (92.1%) of the control nonwords. Although the performance on the network on the regular nonwords is somewhat worse than that of the feedforward networks, it is about equal to the level of performance Glushko (1979) reported for subjects (93.8%; see Table 3). Thus, overall, the ability of the attractor network to pronounce nonwords is comparable to that of skilled readers.

Frequency and Consistency Effects

Figure 14 shows the mean latencies of the network in pronouncing words of varying degrees of spelling-sound consistency as a function of frequency. One of the low-frequency exception words from the Taraban and McClelland (1987) list was withheld from this analysis as it is pronounced incorrectly by the network (SPOOK ⇒/spuk/). Among the remaining words, there are significant main effects of frequency ($F_{3,183}$=25.0, p<.001) and consistency ($F_{3,183}$=8.21, p<.001), and a significant interaction of frequency and consistency ($F_{3,183}$=3.49, p=.017). These effects also obtain in a comparison of only regular and exception words (frequency: $F_{1,91}$=10.2, p=.002; consistency: $F_{1,91}$=22.0, p<.001; frequency-by-consistency: $F_{1,91}$=9.31, p=.003). Considering each level of consistency separately, the effect of frequency is significant for exception words ($F_{1,45}$=11.9, p=.001) and for ambiguous words ($F_{1,46}$=19.8, p=.001) and marginally significant for regular inconsistent words ($F_{1,46}$=3.51, p=.067). There is no effect of frequency among

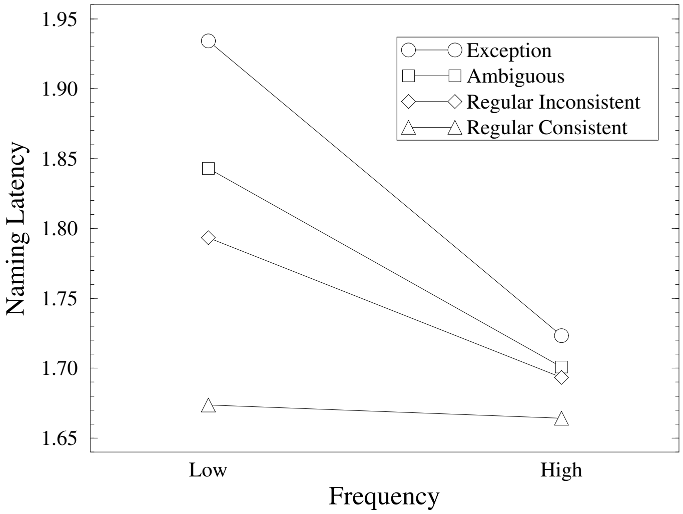

Figure 14: Naming latency of the attractor network trained on actual frequencies for words with various degrees of spelling-sound consistency as a function of frequency.

regular words (F<1).

The naming latencies of the network show a significant effect of consistency for low-frequency words ($F_{3,91}$=6.65, p<.001) but not for high-frequency words ($F_{3,91}$=1.71, p=.170). Among low-frequency words, regular words are significantly different from each of the other three types at p<.05, but regular inconsistent, ambiguous, and exception words are not significantly different from each other (although the comparison between regular inconsistent and exception words is significant at p=.075). Among high-frequency words, none of the pairwise comparisons is significant except between regular and exception words ($F_{1,46}$=4.87, p=.032). Thus, overall, the naming latencies of the network replicate the standard effects of frequency and consistency as found in empirical studies.

Network Analyses

The network's success at word reading demonstrates that, through training, it has developed attractors for the pronunciations of words. How then is it capable of reading nonwords with novel pronunciations? Why isn't the input for a nonword (e.g., MAVE) captured by the attractor for an orthographically similar word (e.g., GAVE, MOVE, MAKE)? We carried out a number of analyses of the network to better understand its ability to read nonwords. Because nonword reading involves recombining knowledge derived from word pronunciation, we were primarily concerned with how separate parts of the input contribute to (a) the correctness of parts of the output, and (b) the hidden representation for the word. As with naming latency, the item SPOOK was withheld from these analyses as it is mispronounced by the network.

Componential Attractors

The first analysis measures the extent to which each phonological cluster (onset, vowel, coda) depends on the input from each orthographic cluster. Specifically, for each word, the activity of the active grapheme units in a particular orthographic cluster were gradually reduced until, when the network was rerun, the phonemes in a particular phonological cluster were no longer correct.[13] This *boundary* activity level measures how important input from a particular orthographic cluster is to the correctness of a particular phonological cluster; a value of 1.0 means that the graphemes in that cluster must be completely active; a value of 0.0 means that the phonemes are completely insensitive to the graphemes in that cluster. In state space, the boundary level corresponds to the radius of the word's attractor basin along a particular direction (assuming state space includes dimensions for the grapheme units).

This procedure was applied to all of the Taraban and McClelland (1987) words as well as to the corresponding set of ambiguous words. Words were excluded from the analysis if they lacked an orthographic onset or coda (e.g., ARE,

[13] Final E was considered to be part of the orthographic vowel cluster.

DO). The resulting boundary values for each combination of orthographic and phonological clusters were subject to an ANOVA with frequency and consistency as between-item factors and orthographic cluster and phonological cluster as within-item factors.

With regard to frequency, high-frequency words have lower boundary values than low-frequency words (0.188 vs. 0.201, respectively; $F_{1,162}$=6.48, p=.012). However, frequency does not interact with consistency ($F_{3,162}$=2.10, p=.102) nor with orthographic or phonological cluster ($F_{2,324}$=1.49, p=.227; and $F_{2,324}$=2.46, p=.087, respectively). Thus, we will consider high- and low-frequency words together in the remainder of the analysis.

There is a strong effect of consistency on the boundary values ($F_{3,162}$=14.5, p<.001), and this effect interacts both with orthographic cluster ($F_{6,324}$=16.1, p<.001) and with phonological cluster ($F_{6,324}$=20.3, p<.001). Figure 15 presents the average boundary values of each orthographic cluster as a function of phonological cluster, separately for words of each level of consistency. Thus, for words of a particular type, the bars for each phonological cluster indicate how sensitive that cluster is to input from each orthographic cluster. Considering regular consistent words first, the figure shows that each phonological cluster depends almost entirely on the corresponding orthographic cluster, and little if at all on the other clusters. For instance, the vowel and coda graphemes can be completely removed without affecting the network's pronunciation of the onset. There is a slight interdependence among the vowel and coda, consistent with the fact that word bodies capture important information in pronunciation (see, e.g., Treiman & Chafetz, 1987). Nonetheless, neither the phonological vowel or coda cluster depend on the orthographic onset cluster. Thus, for a regular word like MUST, an alternative onset (e.g., N) can be substituted and pronounced without depending on or affecting the pronunciation of the body (producing the correct pronunciation of the nonword NUST).

Similarly, for regular inconsistent, ambiguous, and exception words, the correctness of the phonological onset and coda is relatively independent of non-corresponding parts of the orthographic input. The pronunciation of the vowel, however, is increasingly dependent on the orthographic consonants as consistency decreases (main effect of consistency: $F_{3,166}$=47.7, p<.001; p<.05 for all pairwise comparisons). In fact, most spelling-sound inconsistency in English involves unusual vowel pronunciations. Interestingly, for exception words, the vowel pronunciation is less sensitive to the orthographic vowel itself than it is to the surrounding (consonant) context (orthographic onset vs. vowel: $F_{1,41}$=8.39, p=.006; coda vs. vowel: $F_{1,41}$=6.97, p=.012). This makes sense, as the orthographic vowel in an exception word is a misleading indicator of the phonological vowel. Thus, in contrast to regular consistent words, words with ambiguous or exceptional vowel pronunciations depend on the entire orthographic input to be pronounced correctly.

These effects can be understood in terms of the nature of the attractors that develop when training on different types of words. The relative independence of the onset, vowel, and coda correspondences indicates that the attractor basins for regular words consist of three separate, orthogonal sub-basins (one for each cluster). When a word is presented, the network settles into the region in state space where these three sub-basins overlap, corresponding to the word's pronunciation. However, each sub-basin can apply independently, so that "spurious" attractor basins exist where the sub-basins for parts of words overlap (see Figure 16). Each of these combinations corresponds to a pronounceable nonword that the network will pronounce correctly if presented with the appropriate orthographic input. This componentiality arises directly out of the degree to which the network's representations make explicit the structure of the task. By minimizing the extent to which information is replicated, the representations condense the regularities between orthography and phonology. Only small portions of the input and output are relevant to a particular regularity, allowing it to operate independently of other regularities.

The attractor basins for exception words, by contrast, are far less componential than those for regular words (unfortunately, this cannot be depicted adequately in a two-dimensional diagram such as Figure 16). In particular, while the consonant clusters in (most) exception words combine componentially, the correct vowel phoneme depends on the entire orthographic input. Thus, the network develops noncomponential attractor basins for words when necessary. In this way, the network can pronounce exception words and yet still generalize well to nonwords.

The Development of Componentiality in Learning

We can gain insight into the development of this componentiality by returning to the simple, two-layer Hebbian network that formed the basis for the frequency-consistency equation (see Figure 6; also see Van Orden et al., 1990, for related discussion). As expressed by Equation 7, the value of each weight w_{ij} in the network is equal to the sum over training patterns, weighted by the learning rate, of the product of the state of input unit i and the state of output unit j. Patterns for which the input state is 0 do not contribute to the sum, and those for which it is 1 contribute the value of the output state, which is either $+1$ or -1 in this formulation. Thus, the value of the weight can be re-expressed in

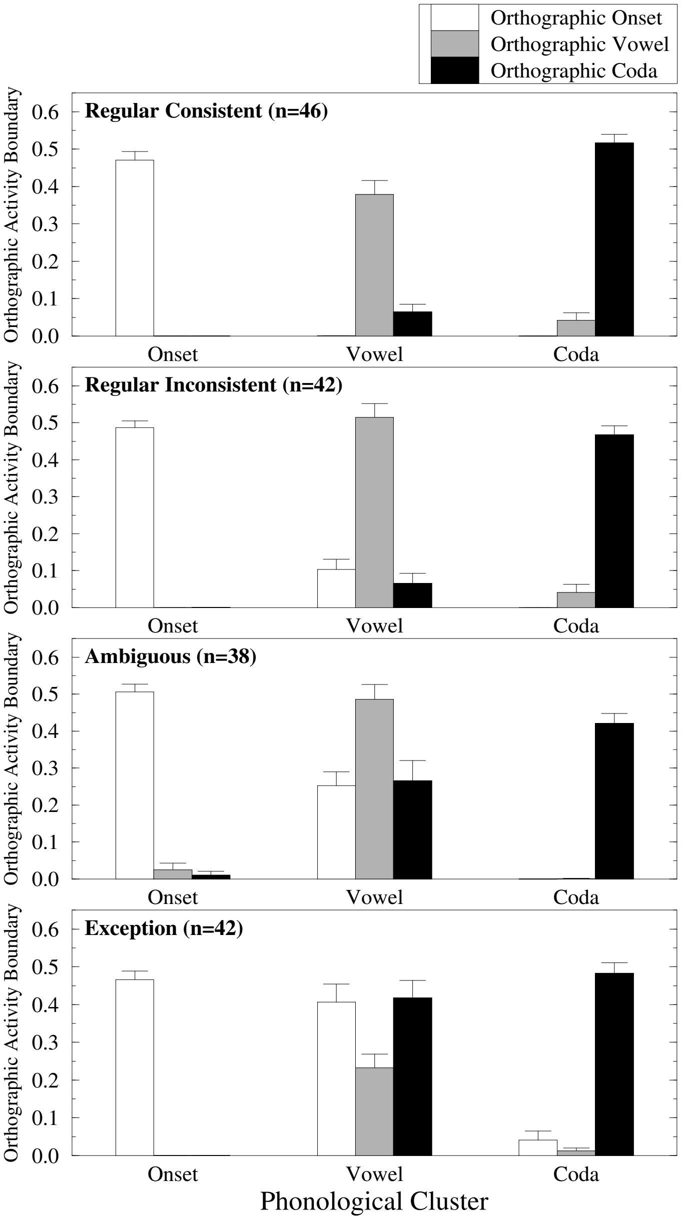

Figure 15: Boundary activity levels of orthographic clusters.

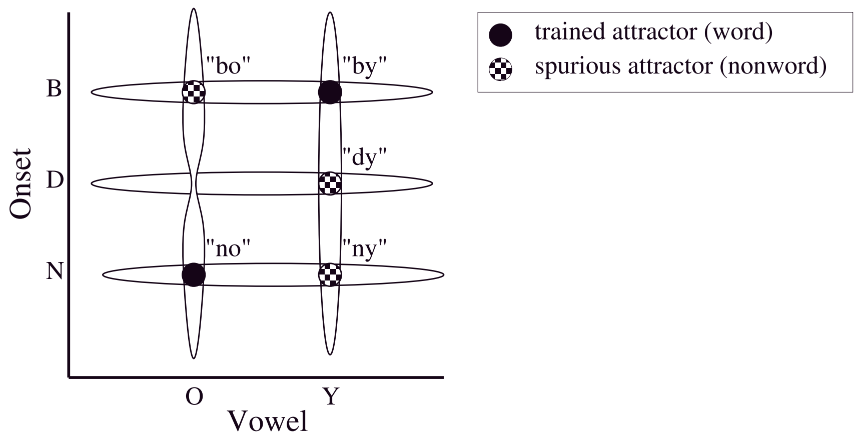

Figure 16: A depiction of how componential attractors for words can recombine to support pronunciations of nonwords. The attractor basins for words consist of orthogonal sub-basins for each of its clusters (only two are depicted here). Spurious attractors for nonwords exist where the sub-basins for parts of words overlap. To support less-componential attractors for exception words (e.g., DO), the sub-basins for vowels in the region of the relevant consonant clusters must be distorted somewhat (into dimensions in state space other than the ones depicted).

terms of two counts: the number of consistent patterns, $N^{[C_{ij}]}$, in which the states of units i and j are both positive, and the number of inconsistent patterns, $N^{[I_{ij}]}$, in which i is positive but j is negative.

$$w_{ij} = \epsilon \left(N^{[C_{ij}]} - N^{[I_{ij}]} \right)$$

If patterns differ in their frequency of occurrence, these counts simply become cumulative frequencies (see Equation 12); for clarity of presentation, we leave this out here (see Reggia et al., 1988, for a simulation based directly on these frequencies).

Now consider a word like PINT ⇒/pInt/. Over the entire set of words, the onset P and /p/ typically co-occur (but not always; cf. PHONE), so that $N^{[C_{ij}]}$ is large and $N^{[I_{ij}]}$ is small, and the weight between these two units becomes strongly positive. By contrast, /p/ never co-occurs with, for example, an onset K (i.e., $N^{[C_{ij}]} = 0$ and $N^{[I_{ij}]}$ is large), leading to a strongly negative weight between them. For onset letters that can co-occur with /p/ and P, such as L, $N^{[C_{ij}]}$ is positive and the resulting weight is thus less negative. Going a step further, /p/ can co-occur with virtually any orthographic vowel and coda, so $N^{[C_{ij}]}$ for each relevant connections is larger and the weight is closer to zero. Actually, given that each phoneme is inactive for most words, its weights from graphemes in non-corresponding clusters will tend to become moderately negative when using Hebbian learning. With error-correcting learning, however, these weights remain near zero because the weights between corresponding clusters are sufficient—and more effective, due to the higher unit correlations—for eliminating the error. These same properties hold for /n/ and /t/ in the coda. Thus, the unit correlations across the entire corpus give rise to a componential pattern of weights for consonant phonemes, with significant values only on connections from units in the corresponding orthographic cluster (see Brousse & Smolensky, 1989, for additional relevant simulations).

The situation is a bit more complicated for vowels. First of all, there is far more variability across words in the pronunciation of vowels as compared with consonants (see Venezky, 1970). Consequently, for connections between vowel graphemes and phonemes, generally $N^{[C_{ij}]}$ is smaller and $N^{[I_{ij}]}$ is larger than for the corresponding onset and coda connections. The more critical issue concerns exceptional vowel pronunciations in words like PINT. Here, for the I—/I/ correspondence, the small $N^{[C_{ij}]}$ is overwhelmed by the large $N^{[I_{ij}]}$ that comes from the much more common I—/i/ correspondence (in which /I/ has a state of -1). Furthermore, with Hebbian learning, the correlations of /i/ with the consonants P, N, and T are two weak to help. Error-correcting learning can compensate to some degree, by allowing the weights from these consonant units to grow larger than dictated by correlation under the pressure to eliminate error. Note that this reduces the componentiality of the vowel phoneme weights. Such cross-cluster weights cannot provide a general solution to pronouncing exception words, however, because, in a diverse corpus, the consonants must be able to co-exist with many other vowel pronunciations (e.g., PUNT, PANT). In order for a network to correctly pronounce exception words while still maintaining the componentiality for regular words (and nonwords), error-correction must be combined with the use of hidden units in order to re-represent the similarities among the words in a way that reduces the interference from inconsistent neighbors (as discussed earlier).

It is important to note, however, that exception words are not noncomponential in a monolithic way, but only in their exceptional aspects. Thus, a word like PINT is three-quarters regular, in the sense that its consonant correspondences contribute to the pronunciations of regular words and nonwords just like those of other items. The traditional dual-route characterization of a lexical "look-up" procedure for exception words fails to do justice to this distinction.

Internal Representations

The first analysis established the componentiality of the attractors for regular words behaviorally, and the second showed how it arises from the nature of learning in a simpler, related system. Hidden units and error correction are required to simultaneously support the noncomponential aspects of word reading, but we have yet to characterize how this is accomplished. The most obvious possibility would be the one raised for the feedforward networks—that the network has partitioned itself into two sub-networks: a componential one for regular words (and nonwords), and a noncomponential one for exception words. As before, however, this does not seem to be the case. If a hidden unit is considered important for pronouncing an item if its removal increases the total error on the item by more than 0.1, then there is a significant positive correlation between the numbers of exception words and the numbers of orthographically-matched nonwords (Taraban & McClelland, 1987) for which hidden units are important (r=.71, t_{98}=9.98, p<.001). Thus, the hidden units have not become specialized for processing particular types of items.

The questions remains, then, as to how the attractor network—as a single mechanism—implements componential attractors for regular words (and nonwords) and noncomponential attractors for exception words. A second analysis attempts to characterize the degree to which hidden representations for regular vs. exception words reflect the differences in the componentiality of their attractors. Specifically, we attempted to determine the extent to which the contribution that an orthographic cluster makes to the hidden representation depends on the context in which it occurs—this should be less for words with more componential representations. For example, consider the onset P in an exception word like PINT. When presented by itself, the onset need only generate its own pronunciation. When presented in the context of _INT, the P must also contribute to altering the vowel from /i/ to /I/. By contrast, in a regular word like PINE, the onset P plays the same role in the context of _INE as when presented in isolation. Thus, if the hidden representations of regular words are more componential than those of exception words, the contribution of an onset (P) should be more greatly affected by the presence of an exception context (_INT) than by a regular context (_INE).

We measured the contribution of an orthographic cluster in a particular context by first computing the hidden representation generated by the cluster with the context (e.g., PINT), and subtracting from this (unit by unit) as a baseline condition, the hidden representation generated by the context alone (e.g., _INT). The contribution of a cluster in isolation was computed similarly, except that the baseline condition in this case is the representation generated by the network when presented with no input (i.e., all grapheme units are set to 0.0). The correlation between these two vector differences was used as a measure of the similarity of the contribution of the cluster in the two conditions. A high correlation indicates that the contribution of a cluster to the hidden representation is independent of the presence of other clusters, and hence, reflects a high degree of componentiality.

These contribution correlations were computed separately for the onset, vowel, and coda clusters of Taraban and McClelland's (1987) frequency-matched regular and exception words. Words lacking either an onset or a coda were withheld from the analysis. The correlations for the remaining words were subject to an ANOVA with frequency and consistency as between-item factors and orthographic cluster as a within-item factor. There was no main effect of frequency ($F_{1,85}$, p=.143) nor any significant interaction of frequency with consistency or orthographic cluster (F<1 for both) so this factor is not considered further. Figure 17 shows the average correlations for regular and exception words as a function of orthographic cluster.

There is no significant interaction of consistency with orthographic cluster (F<1). There is, however, a significant main effect of cluster ($F_{2,170}$=16.1, p<.001), with the vowel cluster producing lower correlations than either consonant cluster (vowel vs. onset: $F1,88$=26.8, p<.001; vowel vs. coda: $F1,88$=21.0, p<.001). More importantly, regular words have higher correlations than exception words [means (standard deviations): 0.828 (0.0506) vs. 0.795 (0.0507), respectively; $F_{1,85}$=20.7, p<.001]. Thus, the contributions of orthographic clusters to the hidden representations are more independent of context in regular words than in exception words. In this sense, the representations of regular words are more componential. What is surprising, however, is that the average correlations for exception words, though

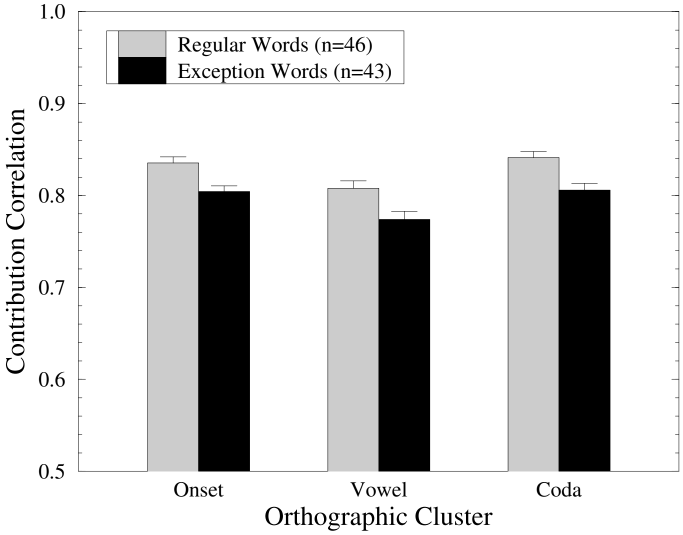

Figure 17: The similarity (correlations) of the contribution that each orthographic cluster makes to the hidden representation in the context of the remaining clusters vs. in isolation, for regular vs. exception words.

lower than those of regular words, are still quite high, and there is considerable overlap between the distributions. Furthermore, the representations for regular words are not completely componential, given that their correlations are significantly less than one.

Apparently, the hidden representations of words only slightly reflect their spelling-sound consistency. An alternative possibility is that these representations capture predominantly *orthographic* information across a range of levels of structure (from individual graphemes to combinations of clusters; cf. Shallice & McCarthy, 1985). If this were the case, the low-order orthographic structure about individual graphemes and clusters could support componential attractors for regular words. The presence of higher-order structure would make the representation of clusters in both regular and exception words somewhat sensitive to context in which they occur. More importantly, this higher-order structure would be particular useful for pronouncing exception words, by overriding at the phonological layer the standard spelling-sound correspondences of individual clusters. In this way, noncomponential attractors for exception words could co-exist with componential attractors for regular words.

To provide evidence bearing on this explanation, a final analysis was carried out to determine the extent to which the hidden representations are organized on the basis of orthographic (as opposed to phonological) similarity. The hidden representations for a set of items are organized orthographically (or phonologically) to the extent that pairs of items with similar hidden representations have similar orthographic (or phonological) representations. Put more generally, the sets representations over two groups of units have the same structure to the extent that they induce the same relative similarities among items.

To control for the contribution of orthography as much as possible, the analysis involved Taraban and McClelland's (1987) body-matched nonwords, regular (inconsistent) words, and exception words (e.g., PHINT, MINT, PINT). For each of these items, we computed the similarity of its hidden representation with the hidden representations of all of the other items (measuring similarity by the correlation of unit activities). The similarities among orthographic representations and among phonological representations were computed analogously. The orthographic, hidden, and phonological similarity values for each item were then correlated in a pairwise fashion (i.e., orthographic-phonological, hidden-orthographic, and hidden-phonological). Figure 18 shows the means of these correlation values for nonwords, regular words, and exception words, as a function of each pair of representation types.

First consider the correlation between the orthographic and phonological similarities. These values reflect the relative amounts of structure in the spelling-sound mappings for different types of items. All of the values are relatively high because of the systematicity of English word pronunciations; even within exception words, the consonant clusters tend to map consistently. Nonetheless, the mappings for exception words are less structured than for nonwords or regular words (paired t_{47}=5.48, p<.001; and t_{47}=5.77, p<.001, respectively). In other words, orthographic similarity is less related to phonological similarity for exception words than for the other items. In a sense, this is the defining characteristic of exception words and, thus, the finding simply verifies that the representations used in the simulations

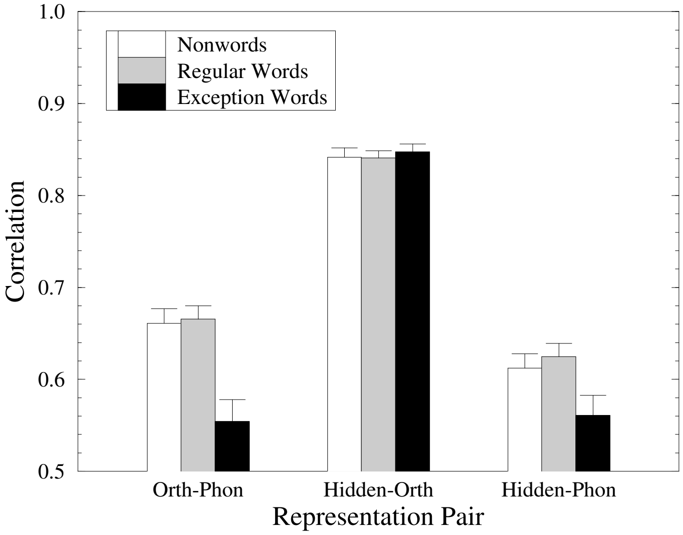

Figure 18: The correlations among orthographic, hidden, and phonological similarities for Taraban and McClelland's (1987) body-matched nonwords, regular (inconsistent) words, and exception words.

have the appropriate similarity structure.

The more interesting comparisons are those that involve the hidden representations. As the Figure shows, across all types of items, the similarities among hidden representations are much more highly correlated with orthographic similarities than with phonological similarities ($p<.001$ for all pairwise comparisons). The representations of nonwords and regular words behave equivalently in this regard. The representations of exception words show the effect even more strongly, having significantly less phonological structure than the other two item types (exception vs. nonword: paired t_{47}=2.81, p=.007; exception vs. regular: paired t_{47}=3.22, p=.002). This may be due to the reliance of these words on high-order orthographic structure to override standard spelling-sound correspondences. Overall, consistent with the explanation offered above, the hidden representations are organized more orthographically than phonologically.

Summary

Interactivity, and its use in implementing attractors, is an important computational principle in connectionist accounts of a wide range of cognitive phenomena. However, the tendency of attractors to capture similar patterns would appear to make them inappropriate for cognitive tasks, such as oral reading, which typically require novel responses to novel inputs (nonwords). The current simulation shows, to the contrary, that using appropriately structured representations leads to the development of attractors with componential structure that supports effective generalization to nonwords. At the same time, the network also develops noncomponential attractors for exception words that violate the regularities in the task. A series of analyses suggest that both componential and noncomponential attractors are supported by hidden representations that represented orthographic information at a range of levels of structure. In this way, attractors provide an effective means of capturing both the regularities and the exceptions in a quasi-regular task.

A further advantage of an attractor network in this domain is that its temporal dynamics in settling to a response provides a more direct analogue of subjects' naming latencies than error in a feedforward network. In fact, the time it takes the network to settle to a stable pronunciation in response to words of varying frequency and consistency replicates the standard pattern found in empirical studies.

Simulation 4: Surface Dyslexia and the Division of Labor Between the Phonological and Semantic Pathways

A central theme of the current work is that the processing of words and nonwords can co-exist within connectionist networks that employ appropriately structured orthographic and phonological representations and operate according to certain computational principles. It must be kept in mind, however, that SM89's general lexical framework—on which the current work is based—contains two pathways by which orthographic information can influence phonological information: a *phonological* pathway and a *semantic* pathway (see Figure 1). Thus far, we have ignored the semantic pathway in order to focus on the principles that govern the operation of the phonological pathway. However, on our view, the phonological and semantic pathways must work together to support normal skilled reading. For example, semantic involvement is clearly necessary for correct pronunciation of homographs like WIND and READ. Furthermore, a semantic variable—imageability—influences the strength of the frequency-by-consistency interaction in the naming latencies and errors of skilled readers (Strain, Patterson, & Seidenberg, 1994). Even in traditional dual-route theories (see, e.g., Coltheart et al., 1993), the lexical procedure must influence the output of the sublexical procedure to account for consistency effects among regular words and nonwords (Glushko, 1979).

The SM89 framework (and the implied computational principles) provides a natural formulation of how contributions from both the semantic and phonological pathways might be integrated in determining the pronunciation of a written word. Critically, when formulated in connectionist terms, this integration has important implications for the nature of *learning* in the two pathways. In most connectionist systems, learning is driven by some measure of the discrepancy or error between the correct response and the one generated by the system. To the extent that the contribution of one pathway reduces the overall error, the other pathway will experience less pressure to learn. As a result, on its own, it may master only those items it finds easiest to learn. Specifically, if the semantic pathway contributes significantly to the pronunciation of words, then the phonological pathway need not master all of the words by itself. Rather, it will tend to learn best those words high in frequency and/or consistency; low-frequency exception words may never be learned completely. This is especially true if there is some intrinsic pressure within the network to prevent overlearning—for example, if weights have a slight bias toward staying small. Of course, the *combination* of the semantic and phonological pathways will be fully competent. But readers of equivalent overt skill may differ in their division of labor between the two pathways (see, e.g., Baron & Strawson, 1976). In fact, as the semantic pathway continues to improve with additional reading experience, the phonological pathway would become increasingly specialized for consistent spelling-sound mappings at the expense of even higher-frequency exception words. At any point, brain damage that reduced or eliminated the semantic pathway would lay bare the latent inadequacies of the phonological pathway. Somewhat paradoxically, those readers with the greatest premorbid competence in reading via meaning would exhibit the most *severe* dyslexia following semantic damage. In this way, a detailed consideration of the division of labor between the phonological and semantic pathways is critical to understanding the specific patterns of impaired and preserved abilities of brain-damaged patients with acquired dyslexia.

Of particular relevance in this context is the finding that brain damage can selectively impair either nonword reading or exception word reading while leaving the other (relatively) intact. Thus, phonological dyslexic patients (Beauvois & Derouesné, 1979) read words (both regular and exception) much better than nonwords, whereas surface dyslexic patients (Marshall & Newcombe, 1973; Patterson et al., 1985) read nonwords much better than (exception) words.

Phonological dyslexia has a natural interpretation within the SM89 framework in terms of selective damage to the phonological pathway (or perhaps within phonology itself; see Patterson & Marcel, 1992), so that reading is accomplished primarily (perhaps even exclusively in some patients) by the semantic pathway. This pathway can pronounce words but is unlikely to provide much useful support in pronouncing nonwords as, by definition, these items have no semantics. Along these lines, as mentioned in the previous section, Plaut and Shallice (1993, also see Hinton & Shallice, 1991) used a series of implementations of the semantic route to provide a comprehensive account of deep dyslexia (Coltheart et al., 1980; Marshall & Newcombe, 1966), a form of acquired dyslexia similar to phonological dyslexia but also involving the production of semantic errors (see Friedman, in press; Glosser & Friedman, 1990, for arguments that deep dyslexia is simply the most severe form of phonological dyslexia). The question of the exact nature of the impairment that gives rise to reading via semantics in phonological dyslexia, and whether this interpretation can account for all of the relevant findings, is taken up in the General Discussion.

Surface dyslexia, on the other hand, would seem to involve reading primarily via the phonological pathway due to an impairment of the semantic pathway. It its purest, *fluent* form (e.g., MP, Behrmann & Bub, 1992; Bub, Cancelliere, & Kertesz, 1985; KT, McCarthy & Warrington, 1986; HTR, Shallice et al., 1983), patients pronounce regular words and nonwords with normal accuracy and latency, but often mispronounce exception words, particularly those of low frequency, by giving a more regular pronunciation (e.g., SEW ⇒ "sue"). Thus, there is a frequency-by-consistency interaction in accuracy that mirrors the interaction in latency exhibited by normal skilled readers (Andrews, 1982; Seidenberg, 1985; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). The relevance of the semantic impairment in surface dyslexia is supported by the finding that, in some cases of semantic dementia (Graham, Hodges, & Patterson, 1994; Patterson & Hodges, 1992; Schwartz, Marin, & Saffran, 1979) and of Alzheimer's type dementia (Patterson, Graham, & Hodges, 1994), the surface dyslexic reading pattern emerges as lexical semantics progressively deteriorates.

The previous simulations of the phonological pathway, along with that of SM89, are similar to surface dyslexic patients in that they read without the aid of semantics. However, the correspondence cannot be direct as, unlike the patients, all of the simulations read exception words as well as skilled readers. One possibility is that surface dyslexia arises from partial impairment of the phonological pathway in addition to severe impairment of the semantic pathway. A more interesting possibility, based on the division-of-labor ideas above, is that the development and operation of the phonological pathway is shaped in an important way by the concurrent development of the semantic pathway, and that surface dyslexia arises when the *intact* phonological pathway operates in isolation due to an impairment of the semantic pathway.

Two sets of simulations are employed to test the adequacy of these two accounts of surface dyslexia. The first set investigates the effects of damage in the attractor network developed in previous simulation. The second involves a new network trained in the context of support from semantics.

Phonological Pathway Lesions

Patterson et al. (1990) investigated the possibility that surface dyslexia might arise from damage to an isolated phonological pathway. They lesioned the SM89 model, by removing different proportions of units or connections, and measured its performance on regular and exception words of various frequencies. The damaged network's pronunciation of a given word was compared with the correct pronunciation and with a plausible alternative—for exception words, this was the regularized pronunciation. Patterson and colleagues found that, after damage, regular and exception words produce about equal amounts of error, and there was no effect of frequency in reading exception words. Exception words were much more likely than regular words to produce the alternative pronunciation, but a comparison of the phonemic features in errors revealed that the network showed no greater tendency to produce regularizations than other errors that differ from the correct pronunciation by the same number of features. Thus, the damaged network failed to show the frequency-by-consistency interaction and the high proportion of regularization errors on exception words characteristic of surface dyslexia.

Using a more detailed procedure for analyzing responses, Patterson (1990) found that removing 20% of the hidden units produced better performance on regular vs. exception words and a (nonsignificant) trend towards a frequency-by-consistency interaction. Figure 19 shows analogous data from 100 instances of lesions to a replication of the SM89 network, in which each hidden unit had a probability p of either 0.2 or 0.4 of being removed. Plotted for each severity of damage is the percent correct on high- and low-frequency regular and exception words (Taraban & McClelland, 1987), the percent of errors on the exception words that are regularizations, and the percent correct on nonwords (Glushko, 1979), counting as correct any pronunciation consistent with that of some word with the same body in the training corpus. Also shown in the figure are the corresponding data for two surface dyslexic patients: MP (Behrmann & Bub, 1992; Bub et al., 1985) and KT (McCarthy & Warrington, 1986).

The milder lesions ($p = 0.2$) produce a good match to MP's performance on the Taraban and McClelland words. However, the more severe lesions ($p = 0.4$) fail to simulate the more dramatic effects shown by KT. Instead, while the damaged network and KT perform about equally well on the high-frequency exception words, the network is not as impaired on the low-frequency exception words and is much more impaired on both high- and low-frequency regular words. In addition, with the less severe damage, only about a third of the network's errors to exception words are regularizations and only just above half of the nonwords are pronounced correctly; for more severe damage, these figures are even lower. By contrast, both MP and KT produce regularization rates around 85-90% and are near perfect at nonword reading. Overall, the attempts to account for surface dyslexia by damaging the SM89 model have been less than satisfactory (see Behrmann & Bub, 1992; Coltheart et al., 1993, for further criticisms).

One possible explanation of this failing parallels our explanation of the SM89 model's poor nonword reading: it is due to the use of representations that do not make the relevant structure between orthography and phonology sufficiently explicit. In essence, the influence of spelling-sound consistency in the model is too weak. This weakness also seems to be contributing to its inability to simulate surface dyslexia after severe damage: regular word reading, nonword reading, and regularization rates are all too low. This interpretation leads to the possibility that a network trained with more appropriately structured representations would, when damaged, successfully replicate the surface dyslexic reading pattern.

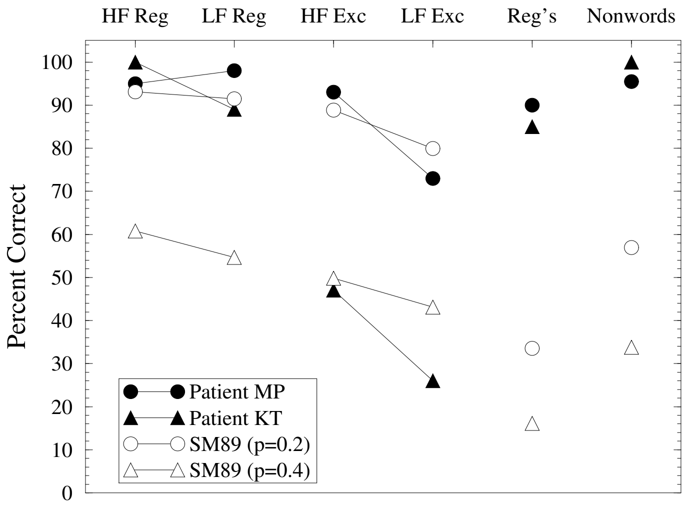

Figure 19: Performance of two surface dyslexic patients (MP, Behrmann & Bub, 1992; Bub et al., 1985; and KT, McCarthy & Warrington, 1986) and of a replication of the SM89 model when lesioned by removing each hidden unit with probability $p = 0.2$ or 0.4 (results are averaged over 100 such lesions). Correct performance is given for Taraban and McClelland's (1987) high- and low-frequency regular and exception words and for Glushko's (1979) nonwords. "Reg's" is the approximate percentage of errors on exception words that are regularizations.

Method

The attractor network was lesioned either by removing each hidden unit or each connection between two groups of units with some probability p, or by adding normally-distributed noise to the weights on connections between two groups of units. In the latter case, the severity of the damage depends on the standard deviation *sd* of the noise—a higher *sd* constitutes a more severe impairment. This form of damage has the advantage over the permanent removal of units or connections of reducing the possibility of idiosyncratic effects from lesions to particular units/connections. As Shallice (1988) has pointed out, such effects in a network simulation are of little interest to the study of the cognitive effects of damage to the brain given the vast difference in scale between the two systems (also see Plaut, in press). In general, simulation studies comparing the effects of adding noise to weights with the effects of removing units or connections (e.g., Hinton & Shallice, 1991) have found that the two procedures yield qualitatively equivalent results.[14]

Fifty instances of each type of lesion of a range of severities were administered to each of the main sets of connections in the attractor network (graphemes-to-hidden, hidden-to-phonemes, phonemes-to-hidden, and phonemes-to-phonemes), and to the hidden units. After a given lesion, the operation of the network when presented with an input and the procedure for determining its response are exactly the same as in Simulation 3.

To evaluate the effects of lesions, the network was tested on Taraban and McClelland's (1987) high- and low-frequency regular and exception words and on Glushko's (1979) nonwords. For the words, in addition to measuring correct performance, we calculated the percentage of errors on the exception words that correspond to a regularized pronunciation. The full list of responses that were accepted as regularizations are given in Appendix 2. As the *undamaged* network mispronounces the word SPOOK, this item was not included in the calculation of regularization rates. For the nonwords, a pronunciation was accepted as correct if it was consistent with the pronunciation of some word in the training corpus with the same body (see Appendix 1).

[14] To see why this should be the case, imagine a much larger network in which the role of each weight in a smaller network is accomplished by the collective influence of a large set of weights. For instance, we might replace each connection in the small network by a set of connections whose weights are both positive and negative and sum to the weight of the original connection. Randomly removing some proportion of the connections in the large network will shift the mean of each set of weights; this will have the same effect as adding a random amount of noise to the value of the corresponding weight in the small network.

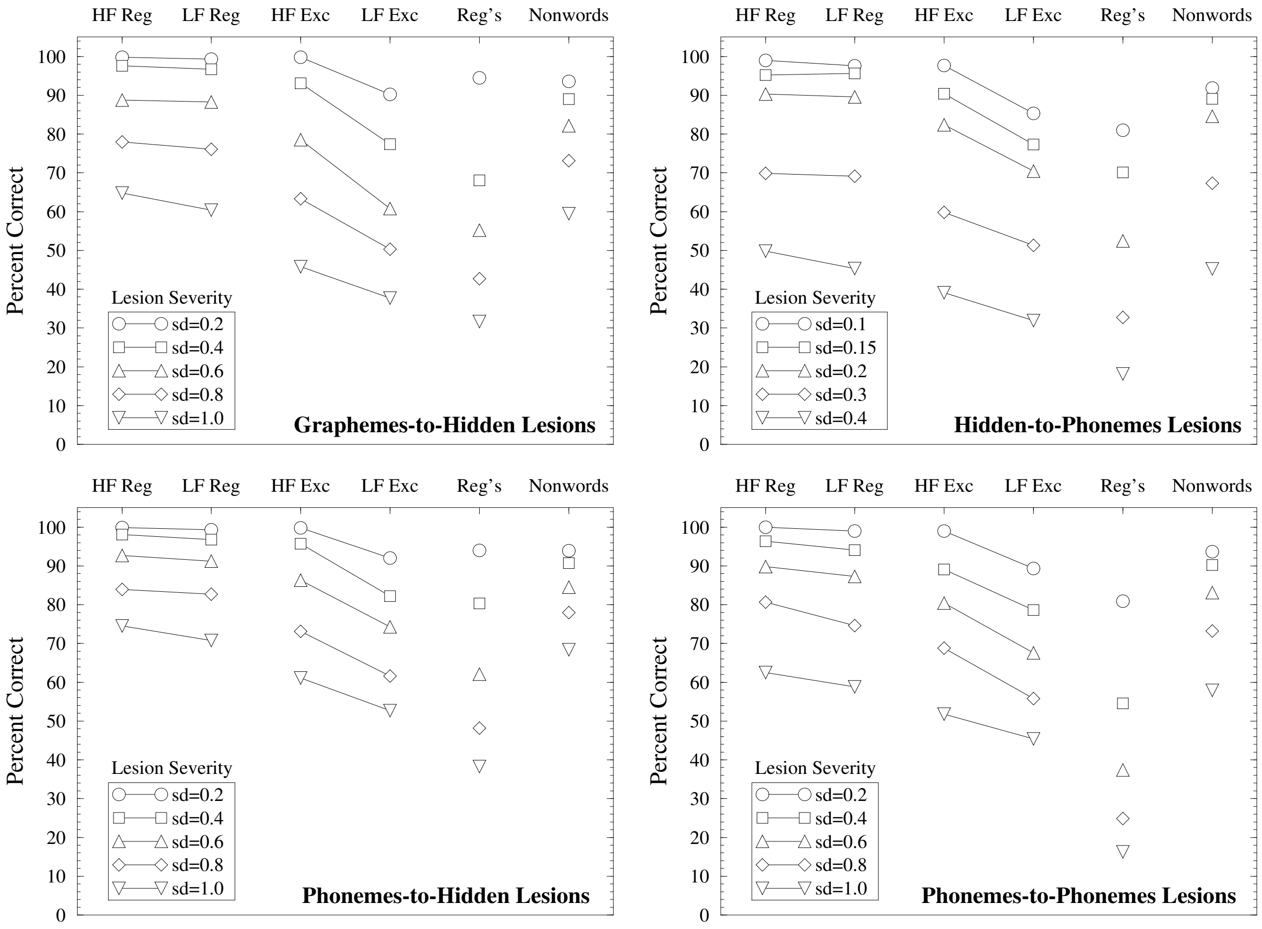

Figure 20: Performance of the attractor network after lesions of various severities to each of the main sets of connections, in which weights are corrupted by noise with mean zero and standard deviation "sd" as indicated. Correct performance is given for Taraban and McClelland's (1987) high- and low-frequency regular and exception words and for Glushko's (1979) nonwords. "Reg's" is the percentage of errors on exception words that are regularizations.

Results and Discussion

Figure 20 shows the data from the attractor network after the weights of each of the four main sets of connections were corrupted by noise of varying severities. The milder lesions to the graphemes-to-hidden connections (on the left of the Figure) produce clear interactions of frequency and consistency in correct performance on word reading. For instance, after adding noise with *sd*=0.4, the network pronounces correctly over 96% of regular words and 93% of high-frequency exception words, but only 77% of low-frequency exception words. In addition, for these lesions, 68% of errors on exception words are regularizations, and 89% of the nonwords are pronounced correctly. Compared with the results from lesions of 20% of the hidden units in the SM89 network, these show a stronger effect of consistency and are a better match to the performance of MP (although the regularization rate is somewhat low; see Figure 19). Thus, as predicted, the use of representations that better capture spelling-sound structure produces a stronger frequency-by-consistency interaction, more regularizations, and better nonword reading.

As found for the SM89 network, however, more severe lesions do not replicate the pattern shown by KT. Lesions that reduce correct performance on high-frequency exception words to equivalent levels (*sd*=1.0; network: 46%; KT: 47%) do not impair low-frequency exception words sufficiently (network: 38%; KT: 26%) and, unlike KT, impair both high- and low-frequency regular words (network: 65% and 60%; KT: 100% and 89%, respectively). Furthermore, and even more unlike KT, there is a substantial drop in both the regularization rate (network: 32%; KT: 85%) and in performance on nonwords (network: 60%; KT: 100%).

Lesions to the other sets of connections produce broadly similar but even weaker results: the frequency-by-

Table 9: Performance of the Attractor Network after Lesions of Units or Connections

	Correct Performance					
	HF Reg	LF Reg	HF Exc	LF Exc	Reg's	Nonwords
Patient MP[a]	95	98	93	73	90[c]	95.5
Patient KT[b]	100	89	47	26	85[c]	100
Attractor Network Lesions						
Graphemes-to-Hidden						
$p = .05$	95.8	94.4	88.9	75.8	65.6	89.6
$p = .3$	49.0	42.8	37.8	27.9	26.0	45.3
Hidden Units						
$p = .075$	93.9	93.5	85.6	75.8	51.4	85.6
$p = .3$	54.5	49.4	45.3	31.7	18.0	48.4
Hidden-to-Phonemes						
$p = .02$	89.0	89.2	81.0	70.0	48.3	82.4
$p = .1$	36.3	31.8	26.4	24.8	13.3	35.5

Note: p is the probability that each of the specified units or connections is removed from the network for a lesion; results are averaged over 50 instances of such lesions.

[a] From Bub et al. (1985) and Behrmann and Bub (1992).

[b] From Shallice et al. (1983).

[c] Approximate.

consistency interactions are weaker (especially for severe lesions), the impairment of regular words is much more severe, and the regularization rates are much lower (note that a different range of lesion severities was used for the hidden-to-phonemes connections as they are much more sensitive to noise). Thus, in summary, mild grapheme-to-hidden lesions in the attractor network can account for MP's behavior, but more severe lesions cannot reproduce KT's behavior.

These negative findings are not specific to the use of noise in lesioning the network; removing units or connections produces qualitatively equivalent results, except that the regularization rates are even lower. To illustrate this, Table 9 presents data for the two patients and for the attractor network after either mild or severe lesions of the graphemes-to-hidden connections, the hidden units, or the hidden-to-phonemes connections. The levels of severity were chosen to approximate the performance of MP and KT on low-frequency exception words.

In summary, some types of lesion to a network implementation of the phonological pathway may be able to approximate the less-impaired pattern of performance shown by MP, but are unable to account for the more dramatic pattern of results shown by KT. These findings suggest that impairment to the phonological pathway may play a role in the behavior of some surface dyslexic patients, but seems unlike to provide a complete explanation of some patients—particular those with normal nonword reading and severely impaired exception word reading.

Phonological and Semantic Division of Labor

We now consider an alternative view of surface dyslexia: that it reflects the behavior of an undamaged but isolated phonological pathway that had learned to depend on support from semantics in normal reading. All of the previous simulations of the phonological pathway have been trained to be fully competent on their own. Thus, if this explanation for surface dyslexia holds, it entails a reappraisal of the relationship between those simulations and the normal skilled word reading system.

The current simulation involves training a new network in the context of an approximation to the contribution of semantics. Including a full implementation of the semantic pathway is, of course, beyond the scope of the present work. Rather, we will characterize the operation of this pathway solely in terms of its influence on the phoneme units within the phonological pathway. Specifically, to the extent that the semantic pathway has learned to derive the meaning and pronunciation of a word, it provides additional input to the phoneme units, pushing them toward their correct activations. Accordingly, we can approximate the influence of the semantic pathway on the development of the phonological pathway by training the latter in the presence of some amount of appropriate external input to the phoneme units.

A difficult issue arises immediately in the context of this approach, concerning the time course of the semantic contribution during the training of the phonological pathway. Presumably, the mapping between semantics and phonology develops, in large part, prior to reading acquisition, as part of speech comprehension and production. By contrast, the orthography-to-semantics mapping, like orthography-to-phonology mapping, obviously can develop only when learning to read. In fact, it is likely that the semantic pathway makes a substantial contribution to oral reading only once the phonological pathway has developed to some degree—in part because of the phonological nature of typical reading instruction, and in part because the orthography-to-phonology mapping is far more structured than the orthography-to-semantics mapping. Also, the degree of learning within the semantic pathway is likely to be sensitive to the frequency with which words are encountered. Accordingly, as a coarse approximation, we will assume that the strength of the semantic contribution to phonology in reading increases gradually over time and is stronger for high-frequency words.

It must be acknowledged that this characterization of semantics fails to capture a number of properties of the actual word reading system that are certainly important in some contexts: differences among words and among phonemes within a word in the contribution of semantics to phonology, interactivity between phonology and semantics, and the relative time course of the semantic and phonological pathways. Nonetheless, the manipulation of external input to the phoneme units allows us to investigate the central claim in the proposed explanation of surface dyslexia: that partial semantic support for word pronunciations alleviates the need for the phonological pathway to master all words, such that, when the support is eliminated by brain damage, the surface dyslexic reading pattern emerges.

Method

As will become apparent below, the necessary simulation requires 4–5 times more training epochs than the corresponding previous simulation. Thus, an attractor network trained on actual word frequencies could not be developed due to the limitations of available computational resources. Rather, the simulation involved training a feedforward network using a square-root compression of word frequencies. Such a network produces a pattern of results in word and nonword reading that is quite similar to the attractor network (see Simulation 2). More importantly, there is nothing specific about the feedforward nature of the network that is necessary to produce the results reported below; an attractor network trained under analogous conditions would be expected to produce qualitatively equivalent results.

The network was trained with the same learning parameters as the corresponding network from Simulation 2 except for one change: a small amount of *weight decay* was introduced, such that each weight experiences a slight pressure to decay towards zero, proportional (with constant 0.001) to its current magnitude. As mentioned above, this provides a bias towards small weights that prevents the network from overlearning and thereby encourages good generalization (see Hinton, 1989). As is demonstrated below, the introduction of weight decay does not alter the ability of the network to replicate the patterns of normal skilled performance on words and nonwords.

Over the course of training, the magnitude S of the input to phoneme units from the (putative) semantic pathway for a given word was set to be

$$S = g \frac{\log(f+2)\,t}{\log(f+2)\,t + k} \tag{16}$$

where f is the Kucera and Francis (1967) frequency of the word, and t is the training epoch. The parameters g and k determine the asymptotic level of input and the time to asymptote, respectively. Their values ($g = 5$, $k = 2000$ in the current simulation), and, more generally, the specific analytic function used to approximate the development of the semantic pathway, affect the quantitative but not the qualitative aspects of the results reported below. Figure 21 shows the mean values of this function over training epochs for the Taraban and McClelland (1987) high- and low-frequency words. If, for a given word, the correct state of a phoneme unit was 1.0, then its external input was positive; otherwise it was the same magnitude but negative.

For the purposes of comparison, a second version of the network was trained without semantics, using exactly the same learning parameters and initial random weights.

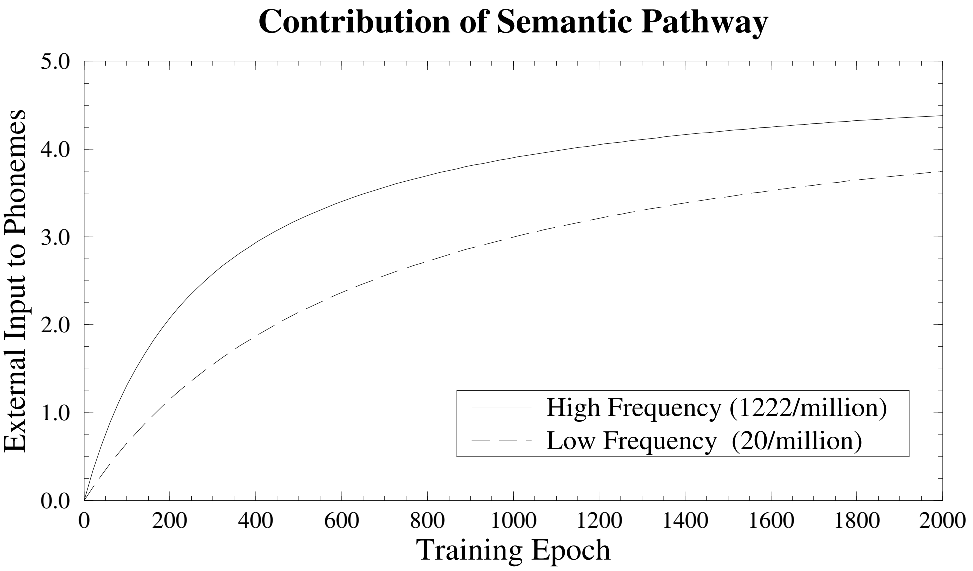

Figure 21: The magnitude of the additional external input supplied to phoneme units by the putative semantic pathway as a function of training epoch, for the Taraban and McClelland (1987) high- and low-frequency words.

Results and Discussion

Learning in the network trained without semantics reached asymptote by epoch 500, at which point it pronounced correctly all but 9 of the 2998 words in the training corpus (99.7% correct). Figure 22 shows the performance of the network on Taraban and McClelland's (1987) high- and low-frequency regular and exception words, and on Glushko's (1979) nonwords, over the course of training. Performance on regular words and nonwords improves quite rapidly over the first 100 epochs, reaching 97.9% for the words and 96.5% for the nonwords at this point. Performance on high-frequency exception words improves somewhat more slowly. By contrast, performance on the low-frequency exception words improves far more slowly, only becoming perfect at epoch 400. At this point, all of the words are read correctly. Even so, there are significant main effects of frequency ($F_{1,92}$=35.9, p<.001) and consistency ($F_{1,92}$=64.3, p<.001), and a significant interaction of frequency and consistency ($F_{1,92}$=26.4, p<.001) in the cross-entropy error produced by the words (means: HFR 0.031, LFR 0.057, HFE 0.120, LFE 0.465). Thus, the network exhibits the standard pattern of normal skilled readers; the introduction of weight decay during training has not substantially altered the basic influences of frequency and consistency in the network.

In the current context, the network trained with a concurrently increasing contribution from semantics (as shown in Figure 21) is the more direct analogue of a normal reader. Not surprising, overall performance improves more rapidly in this case. All of the regulars and the high-frequency exceptions are pronounced correctly by epoch 110, and low-frequency exceptions are at 70.8% correct. By epoch 200, all of the low-frequency exceptions are correct, and nonword reading is 95.4% correct (where we assume nonwords receive no contribution from semantics). At this point, the network with semantics exhibits the standard effects of frequency and consistency in cross-entropy error (means: HFR 0.021, LFR 0.025, HFE 0.102, LFE 0.382; frequency: $F_{1,92}$=19.0, p<.001; consistency: $F_{1,92}$=45.0, p<.001; frequency-by-consistency: $F_{1,92}$=17.8, p<.001). Even after a considerable amount of additional training (epoch 2000), during which the division of labor between the semantic and phonological pathways changes considerably (as shown below), the overt behavior of the normal "combined" network shows the same pattern of effects (nonword reading: 97.7% correct; word cross-entropy error means: HFR 0.013, LFR 0.014, HFE 0.034, LFE 0.053; frequency: $F_{1,92}$=13.6, p<.001; consistency: $F_{1,92}$=125.1, p<.001; frequency-by-consistency: $F_{1,92}$=9.66, p=.003).

This last finding may help explain why, as in previous simulations, networks that are trained to be fully competent on their own replicate the effects of frequency and consistency in naming latency, even though, from the current perspective, such simulations are not fully adequate characterizations of the the isolated phonological pathway in skilled readers. The reason is that, when performance is near asymptote—due either to extended training or to semantic support—word frequency and spelling-sound consistency affect the *relative* effectiveness of processing different words in the same way. This asymptotic behavior follows from the frequency-consistency equation (see

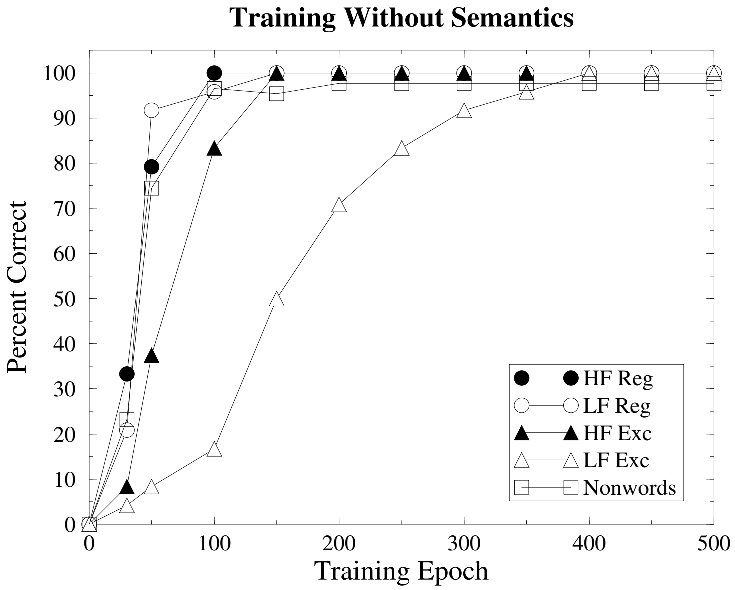

Figure 22: Correct performance of the network trained without semantics on Taraban and McClelland's (1987) high- and low-frequency regular and exception words, and on Glushko's (1979) nonwords, as a function of training epoch.

Equation 12 and Figure 8). Increasing training (by increasing each $N^{[p]}$ in the equation) or adding an additional semantic term to the sum serves equally to drive units further towards their extremal values (also see the General Discussion).

Figure 23 shows the performance of the network at each point in training when the contribution from semantics is eliminated—that is, after a complete semantic "lesion." These data reflect the underlying competence of the phonological pathway when trained in the context of a concurrently developing semantic pathway. First notice that the simulation involves training for 2000 epochs, even thought the bulk of "overt" reading acquisition occurs in the first 100 epochs. Thus, the effects in the network should be thought of as reflecting the gradual improvement of skill from reading experience that, in the human system, spans perhaps many decades.

Initially, performance on nonwords and all types of words improves as the phonological pathway gains competence in the task, much as when the network is trained without semantics (see Figure 22). But as the semantic pathway increases in strength (as characterized by the curves in Figure 21), the accuracy of the combined network's pronunciations of words improves even faster (recall that the combined network is perfect on the Taraban and McClelland words by epoch 200). The pressure to continue to learn in the phonological pathway is thereby diminished. Eventually, at about epoch 400, this pressure is balanced by the bias for weights to remain small. At this point, most of the error that remains comes from low-frequency exception words. This error is reduced as the semantic pathway continues to increase its contribution to the pronunciation of these (and other) words. As a result, the pressure for weights to decay is no longer balanced by the error, and the weights becomes smaller. This causes a *deterioration* in the ability of the phonological pathway to pronounce low-frequency exception words by itself. With further semantic improvement, the processing of high-frequency exception words in the phonological pathway also begins to suffer. Virtually all of the errors on exception words that result from this process are regularizations (plotted as asterisks in the Figure). Larger weights are particularly important for exception words because they must override the standard spelling-sound correspondences that are implemented by many smaller weights. Furthermore, high-frequency words are less susceptible to degradation because any decrement in overt performance induces much large weight changes to compensate. By contrast, the processing of regular words and nonwords is relatively unaffected by the gradual reduction in weight magnitudes. Low-frequency regular words just begin to be affected at epoch 1900.

Thus, with extended reading experience, there is a *redistribution* of labor between the semantic and phonological pathways. As the semantic pathway gains in competence, the phonological pathway increasingly specializes for regular words at the expense of exception words Notice, however, that even with extended training, the phonological pathway continues to be able to read some exception words—particularly those of high-frequency. In this way it is quite unlike the sublexical procedure in a traditional dual-route theory, which can read only regular words and no

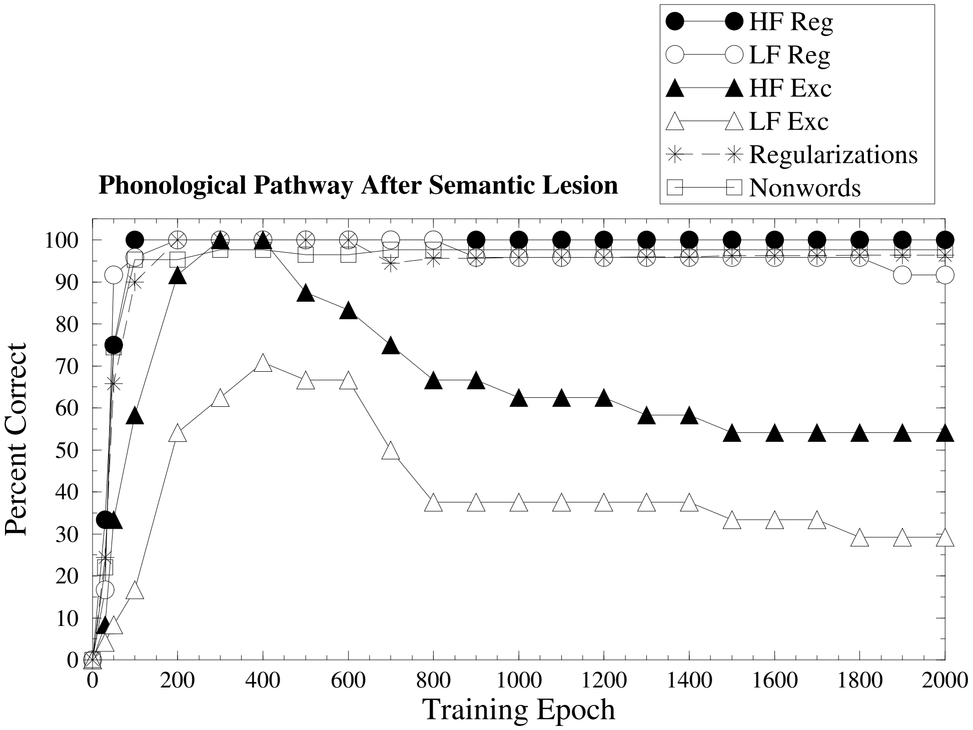

Figure 23: Performance of the network trained with semantics, when semantics is eliminated, on Taraban and McClelland's (1987) high- and low-frequency regular and exception words, and on Glushko's (1979) nonwords, as a function of training epoch.

exception words. It is also important to keep in mind that normal overt performance—as supported by the combination of the phonological and semantic pathways—becomes fully accurate very early on and continues to improve in naming latency (as indirectly reflected by error).

On this interpretation of surface dyslexia, differences among patients in their ability to read exception words may not reflect differences in the severities of their brain damage. Rather, they may reflect differences in their *premorbid* reading experience, with the patients exhibiting the more severe impairment having the greatest premorbid reading competence. To illustrate this more directly, Figure 24 presents data from MP and KT as well as data from the network at two different points in training, when semantics was eliminated. Overall, the network at epoch 400 provides a close match to MP's performance, while the network at epoch 2000 matches KT's performance. The only substantial discrepancy is that, in both conditions, the network's rate of regularizations is higher than that of the corresponding patient (although the patient data are only approximate; see Patterson, 1990). Notice that we are assuming that KT had greater reading experience than MP; while there is no direct evidence on this issue, it is broadly consistent with the observation that KT was a college-educated banker (Shallice et al., 1983), while MP was educated only through high-school (M. Behrmann, personal communication).

Thus far, we have assumed that surface dyslexic patients, at least those of the fluent type, have a lesion that completely eliminates any contribution of the semantic pathway in reading. This assumption may be reasonable for MP and KT, as both patients had very severe impairments in written word comprehension. MP was at chance at selecting which of four written words was semantically related to a given word or picture (also see Bub, Black, Hampson, & Kertesz, 1988 ; Bub et al., 1985. KT's severe word comprehension prevented him from scoring on either the Vocabulary or Similarities subtests of the Wechsler Adult Intelligence Scale (WAIS) (e.g., "bed, bed, I do not know what a bed is;" Shallice et al., 1983, p. 361).

However, some patients with fluent surface dyslexia appear to have only a partial impairment in the semantic pathway. In particular, in patients with semantic dementia, the severity of surface dyslexia appears to be closely related to the severity of semantic impairment (Graham et al., 1994; Patterson & Hodges, 1992). A similar finding applies among patients with Alzheimer's type dementia (Patterson et al., 1994). Such cases have a natural interpretation in the current context in terms of the performance of the network with partial rather than complete elimination of the

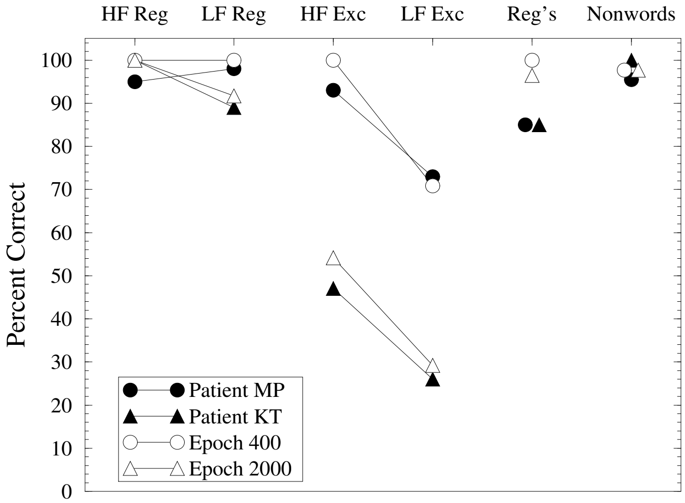

Figure 24: Performance of two surface dyslexic patients (MP, Behrmann & Bub, 1992; Bub et al., 1985; and KT, McCarthy & Warrington, 1986) and the network at different points in training when semantics is eliminated. Correct performance is given for Taraban and McClelland's (1987) high- and low-frequency regular and exception words and for Glushko's (1979) nonwords. "Reg's" is the approximate percentage of errors on exception words that are regularizations.

contribution of the putative semantic pathway. To illustrate this effect, Figure 25 shows the performance of the network trained with semantics to epoch 2000, as the strength of the semantic contribution to the phoneme units—the parameter g in Equation 16—is gradually reduced. As semantics degrades, performance on the low-frequency words is the first to be affected, followed by the high-frequency exception words. By contrast, performance on regular words and nonwords is relatively unaffected by semantic deterioration, although performance on low-frequency regular words is somewhat impaired as semantics is completely eliminated (for $g = 0.0$, the data are identical to those in Figure 23 for epoch 2000). In fact, semantic dementia patients also exhibit a drop in performance on low-frequency regular words when their semantic impairment becomes very severe (Patterson & Hodges, 1992). Of course, a patient with progressive dementia may also have some amount of deterioration within the phonological pathway itself. As Figure 20 and Table 9 illustrate, such impairment would tend to degrade performance on exception words even further, but also would affect performance on regular words and nonwords to some degree.

One final comment with respect to phonological dyslexia seems appropriate. Recall that phonological dyslexic patients are able to pronounce words much better than nonwords. In the current simulation, the external input to the phoneme units that represents the contribution of the semantic pathway is sufficient, on its own, to support accurate word reading (but not nonword reading). On the other hand, severe damage to the phonological pathway certainly impairs nonword reading (see Figure 20 and Table 9). In the limit of a complete lesion between orthography and phonology, nonword reading would be impossible. Thus, a lesion to the network that severely impaired the phonological pathway while leaving the contribution of semantics to phonology (relatively) intact would replicate the basic characteristics of phonological dyslexia.

Summary

The detailed patterns of behavior of acquired dyslexic patients provide important constraints on the nature of the normal word reading system. The most relevant patients in the current context are those with (fluent) surface dyslexia, as, like the networks, they would seem to read without the aid of semantics. These patients read nonwords normally, but exhibit a frequency-by-consistency interaction in word reading accuracy, such that low-frequency exception words are particularly error-prone and typically produce regularization errors. Patterson et al. (1990; Patterson, 1990) were relatively unsuccessful in replicating the surface dyslexia reading pattern by damaging the SM89 model. Although the current simulations employ more appropriately structured representations, when damaged, they too fail to produce surface dyslexia—particularly the more severe form exhibited by KT (Shallice et al., 1983). These findings call into question the interpretation of surface dyslexia as arising from a partial impairment of the phonological pathway in

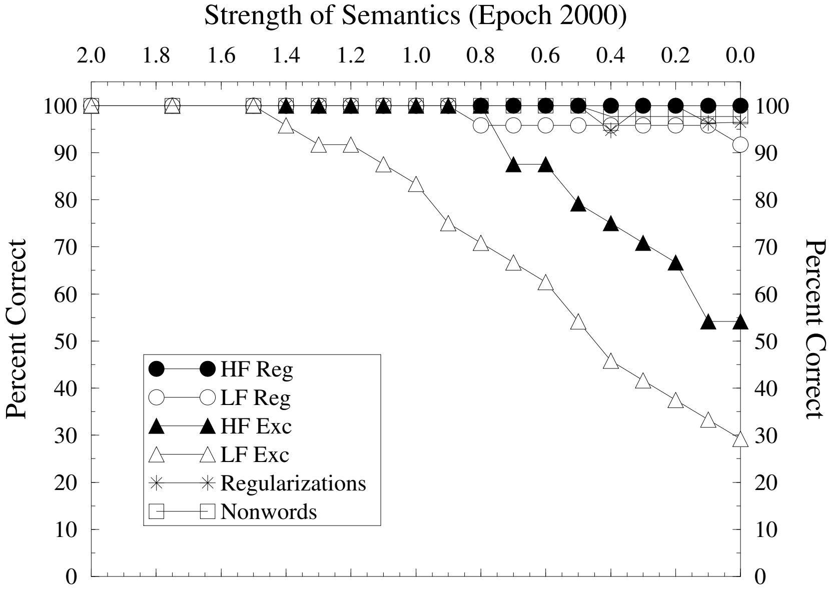

Figure 25: The effect of gradual elimination of semantics on the correct performance of the network after 2000 epochs of training with semantics. Correct performance is given for Taraban and McClelland's (1987) high- and low-frequency regular and exception words, and on Glushko's (1979) nonwords, as a function of training epoch.

addition to extensive impairment of the semantic pathway. Rather, a better match to the surface dyslexic reading pattern—in both its mild and severe forms—is produced by the normal operation of an isolated phonological pathway that develops in the context of support from the semantic pathway. This finding supports a view of the normal word reading system in which there is a division of labor between the phonological and semantic pathways, such that neither pathway alone is completely competent and the two must work together to support skilled word and nonword reading.

General Discussion

The current work develops a connectionist approach to processing in quasi-regular domains, as exemplified by English word reading. The approach derives from a set of general computational principles (McClelland, 1991, 1993); namely, that processing is graded, random, adaptive, interactive, and nonlinear, and that representations and knowledge are distributed. When instantiated in the specific domain of oral reading, these principles lead to a view in which the reading system learns gradually to be sensitive to the statistical structure among orthographic, phonological, and semantic representations, and that these representations simultaneously constrain each other in interpreting a given input.

In support of this view, we have presented a series of connectionist simulations of normal and impaired word reading. A consideration of the shortcomings of a previous implementation (Seidenberg & McClelland, 1989) in reading nonwords led to the development of orthographic and phonological representations that capture better the relevant structure among the written and spoken forms of words. In Simulation 1, a feedforward network employing these representations learned to pronounce all of a large corpus of monosyllabic words, including the exception words, and yet also pronounced nonwords as well as skilled readers.

An analysis of the effects of word frequency and spelling-sound consistency in a related but simpler system formed the basis for understanding the empirical pattern of naming latencies as reflecting an appropriate *balance* between these factors. In Simulation 2, a feedforward network trained with actual word frequencies exhibited good word and nonword reading, and also replicated the frequency-by-consistency interaction in the amount of error it produced for words of various types.

In Simulation 3, a recurrent network replicated the effects of frequency and consistency on naming latency directly in the time required to settle on a stable pronunciation. More critically, the attractors that the network developed for words over the course of training had componential structure that also supported good nonword reading.

Finally, in Simulation 4, the role of the semantic pathway in oral reading was considered in the context of acquired surface dyslexia, in which patients read nonwords well but exhibit a frequency-by-consistency interaction in naming *accuracy*, typically regularizing low-frequency exception words. The view that these symptoms—particularly in their most severe form—reflect the operation of a partially impaired phonological pathway was not supported by the behavior of the attractor network after a variety of types of damage. A further simulation supported an alternative interpretation of surface dyslexia: that it reflects the normal operation of a phonological pathway that is not fully competent on its own because it learned to rely on support from the semantic pathway (which is subsequently impaired by brain damage).

Alternative Perspectives on Word Reading

We can now raise, and then consider in the light of the results summarized above, several issues concerning the nature of the reading process. There is general agreement that (at least) two pathways contribute to reading words and nonwords aloud, but this still leaves open a number of fundamental questions. What are the underlying explanatory principles that determine the existence and the character of these different pathways? How does the operation of each arise from the fundamental principles, and what are the particular principles to which each pathway adheres? How do the different pathways combine to contribute to word and nonword reading? We consider here two very different approaches to these questions.

One view–the so-called dual-route view—holds that the fundamental explanatory principle in the domain of word reading is that distinctly different mechanisms are necessary for reading nonwords on the one hand and exception words on the other. The two mechanisms operate in fundamentally different ways. One assembles pronunciations from phonemes generated by the application of grapheme-phoneme correspondence rules. The other maps whole (orthographic) inputs to whole (phonological) outputs, using either a lexical lookup procedure or, in more recent formulations, an associative network (Pinker, 1991) or McClelland and Rumelhart's (1981) Interactive Activation model (Coltheart et al., 1993).

The alternative view—our connectionist approach—holds that the fundamental explanatory principle in the domain of word reading is that the underlying mechanism employs a nonlinear, similarity-based activation process in conjunction with a frequency-sensitive connection weight adjustment process. Two pathways are necessary in reading, not because different principles apply to items of different types, but because different tasks must be performed. One pathway—here termed the phonological pathway—performs the task of transforming orthographic representations into phonological representations directly. The other—the semantic pathway—actually performs two tasks. The first is specific to reading; namely, the transformation of orthographic representations into semantic representations. The second is a more general aspect of language; namely, the transformation of semantic representations into phonological representations.

At first glance, these two views may appear so similar that deciding between them hardly seems worth the effort. After all, both the lexical procedure in the dual-route account and the semantic pathway in the connectionist account can read words but not nonwords, and both the sublexical procedure and the phonological pathway are critical for nonword reading and work better for regular words than for exception words. It is tempting to conclude that these two explanatory perspectives are converging on essentially the same processing system. Such a conclusion, however, neglects subtle but important differences in the theoretical and empirical consequences of the two approaches.

As a case in point, the sublexical GPC procedure in the dual-route account *cannot* be sensitive to whole-word frequency, as it eschews storage of whole lexical items. By contrast, in the connectionist approach, the phonological pathway maintains an intrinsic and incontrovertible sensitivity to both word frequency and spelling-sound consistency (also see Monsell, 1991). This sensitivity is captured in approximate form by the frequency-consistency equation (Equation 12), which expresses the strength of the response of a simple two-layer network to a given test pattern in terms of the frequency and overlap of the training patterns. The connectionist approach, as reflected by this equation, predicts that there can never be a complete dissociation of frequency and consistency effects; the phonological pathway must always exhibit sensitivity to both. This sensitivity takes a specific form, however: Items that are frequent, consistent or both will have an advantage over items that are neither frequent nor consistent, but items that are frequent and consistent may not enjoy a large additional advantage over those that are only frequent or only consistent; as either frequency or consistency increases, sensitivity to differences in the other decreases.[15]

[15] Recently, Balota and Ferraro (1993) have reported an apparent dissociation of frequency and consistency in the naming latencies of patients with Alzheimer's type dementia, over increasing levels of severity of impairment. However, these patients make substantial numbers of errors, and

This relationship, as we have previously discussed, is approximately characterized by the frequency-consistency equation, which we reproduce here in a form that is elaborated to include a term for the contribution of the semantic pathway, and by separating out the contributions of training patterns whose outputs are consistent with that of the test pattern (i.e., so-called *friends*; McClelland & Rumelhart, 1981) from those whose outputs are inconsistent (i.e., *enemies*). Accordingly, the state $s_j^{[t]}$ of an output (phoneme) unit j that should be on in test pattern t can be written as[16]

$$s_j^{[t]} = \sigma\left(S^{[t]} + \epsilon\left(F^{[t]} + \sum_f F^{[f]} O^{[ft]} - \sum_e F^{[e]} O^{[et]}\right)\right) \tag{17}$$

in which the logistic activation function $\sigma(\cdot)$ is applied to the contribution of the semantic pathway, $S^{[t]}$, plus the contribution of the phonological pathway, which itself is the sum of three terms (scaled by the learning rate, ϵ): (1) the cumulative frequency of training on the pattern itself, (2) the sum of the frequencies of the friends (indexed by *f*) times their overlap with the test pattern, and (3) the sum of the frequencies of the enemies (indexed by e) times their overlap with the test pattern. It must be kept in mind, however, that this equation is only approximate for networks with hidden units and trained by error correction. These two aspects of the implemented networks are critical in that they help to overcome interference from enemies (i.e., the negative terms in Equation 17), thereby enabling the networks to achieve correct performance on exception words—that is, words with many enemies and few if any friends—as well as on regular words and nonwords.

Many of the basic phenomena in the domain of word reading can be seen as natural consequences of adherence to this frequency-consistency equation. In general, any factor that serves to increase the summed input to the activation function, $\sigma(\cdot)$ in Equation 17, improves performance, as measured by naming accuracy and/or latency. Thus, more frequent words are read better (e.g., Forster & Chambers, 1973; Frederiksen & Kroll, 1976) because they have higher values of $F^{[t]}$, and words with greater spelling-sound consistency are read better (Glushko, 1979; Jared et al., 1990) because the positive sum from friends outweighs the negative sum from enemies. The nonlinear, asymptoting nature of the activation function, however, dictates that the contributions of these factors are subject to "diminishing returns" as performance improves. Thus, as reading experience accumulates—thereby increasing $F^{[t]}$, $F^{[f]}$, and $F^{[e]}$ proportionally; or equivalently, increasing ϵ—the absolute magnitudes of the frequency and consistency effects diminish (see, e.g., Backman et al., 1984; Seidenberg, 1985). The same principle applies among different types of stimuli for a reader at a given skill level: performance on stimuli that are strong in one factor is relatively insensitive to variation in other factors. Thus, regular words show little effect of frequency, and high-frequency words show little effect of consistency (as shown in Figure 7). The result is the standard pattern of interaction between frequency and consistency, in which the naming of low-frequency exception words is disproportionately slow or inaccurate (Andrews, 1982; Seidenberg, 1985; Seidenberg et al., 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985).

The elaborated version of the frequency-consistency equation also provides a basis for understanding the effects of semantics on naming performance. In the approximation expressed by Equation 17, the contribution of the semantic pathway for a given word, $S^{[t]}$, is simply another term in the summed input to each output (phoneme) unit. Just as with frequency and consistency, then, a stronger semantic contribution moves the overall input further along the asymptoting activation function, thereby diminishing the effects of other factors. As a result, words with a relatively weak semantic contribution (e.g., abstract or low-imageability words; Jones, 1985; Saffran, Bogyo, Schwartz, & Marin, 1980) exhibit a stronger frequency-by-consistency interaction—in particular, naming latencies and error rates are disproportionately high for items that are weak on all three dimensions: abstract, low-frequency exception words (Strain et al., 1994).

Of course, as the simulations demonstrate, networks with hidden units and trained with error correction can learn to pronounce correctly all types of words without any help from semantics. In the context of the more general framework, however, full competence is required only from the combination of semantic and phonological influences. Thus, as the semantic pathway develops and $S^{[t]}$ increases, the contribution required from the other, phonological terms in Equation 17 to achieve the same level of performance is correspondingly reduced. With the additional assumption that the system has an intrinsic bias against unnecessary complexity (e.g., by limiting its effective degrees of freedom with weight decay), extended reading experience leads to a *redistribution* of labor. Specifically, as the semantic pathway improves, the phonological pathway gradually loses its ability to process the words it learned most weakly: those that

the usual relationship of frequency and consistency holds in their accuracy data (also see Patterson et al., 1994). Furthermore, the dissociation was not found in naming latencies of younger or older normal subjects.

[16] For a unit with a target of -1, the signs would simply be reversed. Alternatively, the equation can be interpreted as reflecting the correlation of the activation of output unit j with its target, which may in that case be either $+1$ or -1.

are low in both frequency and consistency.

If, in this context, the contribution from semantics is severely weakened or eliminated (by brain damage), the summed input to each output unit will be reduced by as much as $S^{[t]}$. For output units with significant negative terms in their summed input—that is, for those in words with many enemies—this manipulation may cause their summed input (and hence their output) to change sign. The result is an incorrect response. Such errors tend to be regularizations because the reduced summed input affects only those output units whose correct activations are inconsistent with those of the word's neighbors. Furthermore, as frequency makes an independent positive contribution to the summed inputs, errors are more likely for low- than for high-frequency exception words. By contrast, a reduction in the contribution from semantics has little if any effect on correct performance on regular words because the positive contribution from their friends is sufficient on its own to give output units the appropriately signed summed input. The resulting pattern of behavior, corresponding to fluent surface dyslexia (Bub et al., 1985; McCarthy & Warrington, 1986; Shallice et al., 1983), can thus be seen as an exaggerated manifestation of the same influences of frequency and consistency that give rise to the normal pattern of naming latencies.

The pattern of joint, nonlinear sensitivity to the combined effects of frequency and consistency in the connectionist account, along with assumptions on the contribution of semantics, lead to a number of predictions not shared by traditional dual-route accounts. First, frequency and consistency can trade off against each other, so that the detrimental effects of spelling-sound inconsistency can always be overcome by sufficiently high word frequency. Consequently, the connectionist account makes a strong prediction: there *cannot* be an (English-language) surface dyslexic patient who reads *no* exception words; if regular words can be read normally, there must also be some sparing of performance on high-frequency exceptions. By contrast, a dual-route framework could account for such a patient quite easily, in terms of damage that eliminates the lexical route(s) while leaving the GPC route in operation. In fact, given the putative separation of these routes, the framework would seem to *predict* the existence of such patients. The connectionist account also differs from the dual-route account in claiming that consistency rather than regularity *per se* (i.e., adherence to GPC rules) is the determining variable in "regularization" errors (where, as formulated here, consistency depends on all types of orthographic overlap rather than solely on word bodies; cf. Glushko, 1979). Finally, the connectionist account predicts a close relationship between impairments to semantics and the surface dyslexic reading pattern (Graham et al., 1994; Patterson & Hodges, 1992; although, for readers with well-developed phonological pathways—perhaps due to a weak semantic pathway—the pattern may emerge only with very severe semantic impairment; see, e.g., Schwartz, Saffran, & Marin, 1980). By contrast, dual-route theories that include a lexical, non-semantic pathway (e.g., Coltheart, 1978, 1985; Coltheart et al., 1993) predict that selective semantic damage should have no effect on naming accuracy.

Our connectionist account, we believe, also has an important advantage of simplicity over the dual-route approach. This advantage goes well beyond the basic point that it provides a single mechanism that can account for exception word and nonword reading, while the dual-route model must rely on two mechanisms. The additional advantage lies in the fact that the boundary between regular and exception words is not clear, and all attempts to draw such a boundaries lead to unfortunate consequences. First, the marking of items as exceptions which must be looked up as wholes in the lexicon ignores the fact that most of the letters in these items will take their standard grapheme-phoneme correspondences. Thus, in PINT, 3/4 of the letters take their regular correspondence. Second, the marking of such items as exceptions ignores the fact that even the parts that are exceptional admit of some regularity, so that, for example, vowels generally still correspond to phonemes that they correspond to in many other words. Thus, in PINT, the I corresponds to the same vowel that it corresponds to most words ending in I_E (where the "_" represents any consonant), as well as _IND and _ILD. Third, exceptions often come in clusters that share the same word-body. Special word-body rules may be invoked to capture these clusters, but then this makes words that conform to the more usual correspondence exceptional. Thus, we could treat OO ⇒/u/ when followed by K as regular, but this would make SPOOK, which takes the more common correspondence OO ⇒/U/, an exception. The placement of virtually any word into the exception category, then, neglects its partial regularity and prevents the word both from benefitting from this partial regularity and from contributing to patterns of consistency it enters into with other items. Our connectionist approach, by contrast, avoids the need to impose such unfortunate divisions, and leaves a mechanism that exhibits sensitivity to all these partially regular aspects of so-called exception words.

The fact that exceptions are subject to the same processes as all other items in our system allows us to explain why there are virtually no completely arbitrary exceptions. On the other hand, the dual-route approach leaves this fact of the spelling-sound system completely unexplained. Nor, in fact, do some dual-route models even provide a basis for accounting for effects of consistency in reading words and nonwords. Recent dual-route theorists (e.g., Coltheart et al., 1993) have appealed to partial activation of other lexical items as a basis for such effects. Such an assumption moves part-way toward our view that consistency effects arise from the influence of all lexical items. We would only add that our connectionist model exhibits these effects as well as the requisite sensitivity to general grapheme-phoneme correspondences, without stipulating a separate rule system over and above the system that exhibits the broad range of consistency effects.

Additional Empirical Issues

Proponents of dual-route theories have raised a number of empirical issues that they believe challenge our connectionist account of normal and impaired word reading. For example, Coltheart et al. (1993, also see Besner et al., 1990) raise six questions concerning the reading process, all but one of which—exception word reading—they deem problematic for the SM89 framework. Two of the remaining five—nonword reading and acquired surface dyslexia—have been addressed extensively in the current work. Here we discuss how the remaining three issues—acquired phonological dyslexia, developmental dyslexia, and lexical decision—may be accounted for in light of these findings. We also consider two other empirical issues that have been interpreted as providing evidence against the current approach—pseudohomophone effects (Buchanan & Besner, 1993; Fera & Besner, 1992; McCann & Besner, 1987; Pugh, Rexer, & Katz, 1994) and blocking effects (Baluch & Besner, 1991; Monsell et al., 1992).

Acquired Phonological Dyslexia

As mentioned earlier, it is straightforward within the SM89 framework to account for the central characteristic of acquired phonological dyslexia—substantially better word reading than nonword reading—in terms of a relatively selective impairment of the phonological pathway. The apparent difficult arises when considering patients who (a) are virtually unable to read nonwords, suggesting a complete elimination of the phonological pathway, and (b) have an additional semantic impairment that seems to render the semantic pathway insufficient to account for the observed proficiency at word reading. Two such patients have been described in the literature: WB (Funnell, 1983) and WT (Coslett, 1991). To explain the word reading of these patients, dual-route theorists claim that it is necessary to introduce a third route that is lexical but nonsemantic.

In point of fact, Coltheart et al. (1993) explicitly considered an alternative explanation and (we think too hastily) rejected it.

> Perhaps a patient with an impaired semantic system, who therefore makes semantic errors in reading comprehension and who also has a severely impaired nonsemantic reading system, could avoid making semantic errors in reading aloud by making use of even very poor information about the pronunciation of a word yielded by the nonsemantic reading system. The semantic system may no longer be able to distinguish the concept *orange* from the concept *lemon*; however, to avoid semantic errors in reading aloud, all the nonsemantic route needs to deliver is just the first phoneme of the written word, not a complete representation of its phonology. (p. 596)

Coltheart and colleagues argued against this account entirely on the basis of two findings of Funnell (1983): WB did not pronounce correctly any of a single list of 20 written nonwords, and he did not give the correct phonemic correspondence to any of 12 single printed letters. Thus, they claimed, "WB's nonsemantic reading route was not just severely impaired, it was completely abolished" (p. 596).

This argument is unconvincing. First of all, it would seem unwise to base such a strong theoretical claim on so few empirical observations, especially given how little information is required of the phonological pathway on the above account. To pronounce a nonword correctly, however, *all* of its phonemes must be derived accurately. Thus, WB's inability to read 20 nonwords cannot be taken as definitive evidence that his phonological pathway is completely inoperative. Furthermore, WB did, in fact, make semantic errors in oral reading (e.g., TRAIN ⇒"plane", GIRL ⇒"boy"; see Funnell, 1983, Appendix 1). Although such errors were relatively rare, comprising only 7.5% (5/67) of all lexical error responses, there were *no* error responses that were completely unrelated to the stimulus. Thus, the effect of semantic relatedness in errors is difficult to ascribe to chance responding (see Ellis & Marshall, 1978; Shallice & McGill, 1978). More generally, fully 38.8% (26/67) of WB's lexical errors had a semantic component, typically in combination with visual/phonemic or morphological relatedness.

More critically, Coltheart and colleagues fail to take into account the fact that WB exhibited deficits on purely phonological tasks, such as nonword repetition (Funnell, 1983) and phoneme stripping and blending (Patterson & Marcel, 1992), suggesting an additional impairment within phonology itself. Funnell had argued that such a phonological impairment could not explain WB's nonword reading deficit, because (a) he repeated nonwords more successfully (10/20) than he read them (0/20), and (b) he achieved some success (6/10) in blending three-phoneme words from auditory presentation of their individual phonemes. We note, however, that the failure to repeat fully half of a set of simple, single-syllable, word-like nonwords (e.g., COBE, NUST) certainly represents a prominent phonological deficit. Moreover, since Funnell's auditory blending test used only words as target responses, WB's partial success on this task is not especially germane to the issue. Patterson and Marcel (1992) assessed WB's blending performance with nonword targets and found that he was unable to produce a single correct response, whether the auditory presentation consisted of the three individual phonemes of a simple nonword (such as COBE) or its onset and rime. Patterson and Marcel argued that this phonological deficit in a *non-reading* task was sufficient to account for WB's complete inability to read nonwords.

Thus, the pattern of performance exhibited by WB can be explained within the SM89 framework in terms of a mildly impaired semantic reading pathway, an impaired phonological reading pathway and, in particular, an impairment within phonology itself. A similar explanation applies to WT (Coslett, 1991): although this patient's performance on phonological blending tasks is not reported, she was severely and *equally* impaired in her ability to read and to repeat the same set of 48 nonwords.

We point out in passing that *deep* dyslexia (Coltheart et al., 1980), the remaining major type of acquired central dyslexia and closely related to phonological dyslexia (see, e.g., Glosser & Friedman, 1990), can be accounted for in terms of the same computational principles that are employed in the current work (see Plaut & Shallice, 1993).

Developmental Dyslexia

Our focus in the current work has been on characterizing the computational principles governing normal skilled reading and acquired dyslexia following brain damage in premorbidly literate adults. Even so, we believe that the same principles provide insight into the nature of reading acquisition, both in its normal form and in developmental dyslexia, in which children fail to acquire age-appropriate reading skills.

There is general agreement that a number of distinct patterns of developmental dyslexia exist, although exactly what these patterns are and what gives rise to them is a matter of ongoing debate. A common viewpoint is that there are developmental analogues to the acquired forms of dyslexia (see, e.g., Baddeley, Ellis, Miles, & Lewis, 1982; Harris & Coltheart, 1986; Marshall, 1984). Perhaps the clearest evidence comes from Castles and Coltheart (1993), who compared 53 dyslexic children with 56 age-matched normal readers in their ability to pronounce exception words and nonwords. The majority (32) of the dyslexic children were abnormally poor on both sets of items. However, 10 were selectively impaired at exception word reading, corresponding to developmental surface dyslexia, and 8 were selectively impaired at nonword reading, corresponding to developmental phonological dyslexia. Castles and Coltheart interpret their findings as supporting a dual-route theory of word reading, in which either the lexical or the sublexical procedure can selectively fail to develop properly (although they offer no suggestion as to why this might be).

More recently, Manis, Seidenberg, Doi, McBride-Chang, and Peterson (submitted) compared 51 dyslexic children with 51 controls matched for age and 27 matched for reading level. They confirmed the existence of separate surface and phonological dyslexic patterns although, again, most of the dyslexic children showed a general reading impairment. Critically, the performance of the developmental surface dyslexic children was remarkably similar to that of reading-level matched controls, suggesting a developmental delay. By contrast, the phonological dyslexic children performed unlike either set of controls, suggesting a deviant developmental pattern. While these findings are not incompatible with the dual-route account, Manis and colleagues contend that they are more naturally accounted for in terms of different impediments to the development of a single (phonological) pathway. Specifically, they suggest (following SM89) that the delayed acquisition in developmental surface dyslexia may arise from limitations in the available computational resources within the phonological route. Consistent with this interpretation, SM89 found that a version of their network, trained with only half the normal number of hidden units, showed a disproportionate impairment on exception words compared with regular words (although performance on all items was poorer, consistent with finding that generalized deficits are most common). However, the nonword reading capability of the network was not tested, and Coltheart et al. (1993) point out that it was not likely to be very good, given that overall performance was worse than in the normal network which itself is impaired on nonword reading.

Just as for normal skilled reading, this limitation of the SM89 model stems from its use of inappropriately structured orthographic and phonological representations. To demonstrate this, we trained a feedforward network with only 30 hidden units in an identical fashion to the one with 100 hidden units from Simulation 4 (without semantics). This network was chosen for comparison simply because it is the only one for which the relevant acquisition data has already

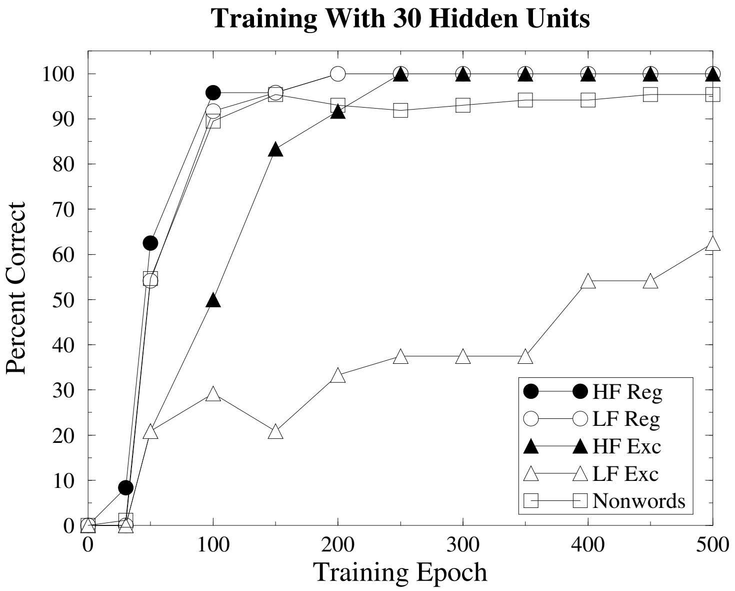

Figure 26: Correct performance of a feedforward network with only 30 hidden units on Taraban and McClelland's (1987) high- and low-frequency regular and exception words, and on Glushko's (1979) nonwords, as a function of training epoch. The network was trained exactly as the one whose corresponding data are shown in Figure 22.

been presented, in Figure 22—the other networks would be expected to show similar effects. The corresponding data for the version with 30 hidden units are given in Figure 26. As a comparison of the figures reveals, limiting the number of hidden units selectively impairs performance on exception words, particularly those of low frequency. By contrast, nonword reading is affected only very slightly. Notice that the performance of the dyslexic network at epoch 500 is quite similar to that of the normal network at about epoch 150. Thus, limiting the computational resources that are available for learning the spelling-to-sound task reproduces the basic delayed pattern of developmental surface dyslexia. Of course, other manipulations that impede learning, such as weak or noisy weight changes, would be expected to yield similar results.

With regard to developmental phonological dyslexia, Manis et al. (submitted) suggest that a selective impairment in nonword reading may arise from the use of phonological representations that are poorly articulated, perhaps due to more peripheral disturbances (also see, e.g., Liberman & Shankweiler, 1985; Rack, Snowling, & Olson, 1992). A consideration of the normal SM89 model is instructive here. That network employed representations that, we have argued, poorly capture the relevant structure within and between orthography and phonology. As a result, the model was over 97% correct at reading words, both regular and exception, but only 75% correct on a subset of Glushko's (1979) nonwords (when scored appropriately; see Seidenberg & McClelland, 1990). Thus, in a sense, the model behaved like a mild phonological dyslexic (see Besner et al., 1990, for similar arguments). In this way, the performance of the model provides evidence that a system with adequate computational resources, but which fails to develop appropriately componential orthographic and (particularly) phonological representations, will also fail to acquire normal proficiency in sublexical spelling-sound translation. It should also be kept in mind that, to whatever extent the semantic pathway develops and contributes during reading acquisition, the dissociation between word and nonword reading would be exacerbated.

A final point of contention with regard to the implications of developmental reading disorders for the SM89 framework concerns the existence of children whose oral reading ability, even on exception words, far surpasses their comprehension—as in so-called *hyperlexia* (Huttenlocher & Huttenlocher, 1973; Mehegan & Dreifuss, 1972; Metsala & Siegel, 1992; Silverberg & Silverberg, 1967). Typically, these children are moderately to severely retarded on standardized intelligence tests, and may totally lack conversational speech. They also tend to devote a considerable amount of time and attention to reading, although this has not been studied thoroughly. We suggest that, perhaps due to abnormally poor development in the semantic pathway, such children may have phonological pathways that are like our networks trained without semantics. In the limit, such networks learn to pronounce all types of words and nonwords accurately with *no* comprehension.

Lexical Decision

The final of Coltheart et al.'s (1993) objections to the SM89 model concerns its ability to perform lexical decisions. While SM89 establish that, under some stimulus conditions, the model can discriminate words from nonwords on the basis of a measure of its accuracy in regenerating the orthographic input, Besner and colleagues (Besner et al., 1990; Fera & Besner, 1992) have demonstrated that its accuracy in doing so is worse than that of human subjects in many conditions. Coltheart et al. (1993) mistakenly claim that the SM89 orthographic error scores yield a false-positive rate of over 80% on Waters and Seidenberg's (1985) nonwords when word error rates are equated with subjects' at 6.1%—in fact, these numbers result from using *phonological* error scores (Besner et al., 1990), which SM89 do not employ (although they do suggest that learning phonological attractors for words might help). While the actual false-positive rate is much lower—Besner and colleagues report a rate of 28% when orthographic and phonological error scores are summed and orthographically strange words are excluded—it is still unsatisfactory.

Of course, SM89 never claimed that orthographic and phonological information are completely sufficient to account for lexical decision performance under all conditions, pointing out that "there may be other cases in which subjects must consult information provided by the computation from orthography to semantics" (p. 552). Semantics is a natural source of information on which to distinguish words from nonwords, given that, in fact, a string of letters or phonemes is *defined* to be a word by virtue of it having a meaning. Coltheart and colleagues raise the concern that, in a full implementation of the SM89 framework, the presentation of an orthographically regular nonword (e.g., SARE) would activate semantics to the same degree as a word (e.g., CARE), thereby precluding lexical decision.

While further simulation work is clearly required to address the full range of lexical decision data adequately, a few comments may serve to allay this specific concern. We imagine that the semantic representations for words are relatively *sparse*, meaning that each word activates very few of the possible semantic features, and each semantic feature participates in the meanings of a very small percentage of words. Connectionist networks of the sort we are investigating learn to set the base activation level of each output unit to the expected value of its correct activations across the entire training corpus, because these values minimize the total error in the absence of any information about the input. In the case of sparse semantic representations, this means that semantic features would be almost completely inactive without specific evidence from the orthographic input that they should be active. Notice that the nature of this evidence must be *very* specific in order to prevent the semantic features of a word like CARE from being activated by the presentation of orthographically similar words like ARE, SCARE, CAR, etc. This extreme sensitivity to small orthographic distinctions would also prevent semantic features from being activated by a nonword like SARE. Thus, on this account, the computational requirements of a connectionist system that maps orthography to semantics veritably *entail* the ability to perform lexical decision.

Pseudohomophone and Blocking Effects

Two other, somewhat overlapping sets of empirical findings have been viewed as problematic for the current approach: pseudohomophone effects (Buchanan & Besner, 1993; Fera & Besner, 1992; McCann & Besner, 1987; Pugh et al., 1994) and blocking effects (Baluch & Besner, 1991; Monsell et al., 1992). The first set involves demonstrations that, under a variety conditions, pseudohomophones (i.e., nonwords with pronunciations that match that of a word; e.g., BRANE) are processed differently than orthographically-matched nonpseudohomophonic nonwords (e.g., FRANE). For example, subjects are faster to name pseudohomophones and slower (and less accurate) to reject them in lexical decision (McCann & Besner, 1987). The second set of problematic findings involves demonstrations that subjects' performance is sensitive to the context in which orthographic stimuli occur, usually operationalized in terms of how stimuli are blocked together during an experiment. For example, subjects are slower and make more regularization errors when pronouncing exception words intermixed with nonwords than when pronouncing pure blocks of exception words (Monsell et al., 1992).

Neither of these sets of phenomena are handled particularly well by the SM89 implementation, but have natural formulations within the more general framework that includes semantics. Pseudohomophone effects may stem from interactions between phonology and semantics that do not occur for control nonwords. Blocking effects may reflect the ability of subjects to adaptively control the relative contribution of the semantic and phonological pathways in lexical tasks. These interpretations are supported by recent findings of Pugh et al. (1994), who investigated effects of spelling-sound consistency and semantic relatedness in lexical decision, as a function of whether or not the nonword foils include pseudohomophones. They found faster latencies for consistent words than for inconsistent words only in the context of purely nonpseudohomophonic nonwords; there was no effect of consistency when pseudohomophones were present. Similarly, in a dual lexical decision paradigm, they obtained facilitation for visually similar word pairs that are phonological consistent (e.g., BRIBE–TRIBE) and inhibition for those that are inconsistent (e.g., COUCH–TOUCH; Meyer et al., 1974) only when no pseudohomophones were present; the introduction of pseudohomophones eliminated the consistency effect. However, semantic relatedness (e.g., OCEAN–WATER) yielded facilitation regardless of nonword context. These findings suggest that subjects normally use both the semantic and phonological pathways in lexical decision, but avoid the use of the phonological pathway when this would lead to inappropriate semantic activity, as when pseudohomophones are included as foils.

Implications for Other Domains

The approach we have taken here is applicable, we believe, to a wide range of linguistic and cognitive domains—essentially, to all domains in which there is quasi-regular structure. The first domain to which the approach was applied was the domain of inflectional morphology (Rumelhart & McClelland, 1986). As stated in the Introduction, this application certainly remains controversial; Pinker and his colleagues (Marcus et al., 1992; Pinker, 1991; Pinker & Prince, 1988) continue to maintain that no single mechanism can fully capture the behavior of the regular inflectional process and the handling of exceptions. While we do not claim that the existing connectionist simulations have fully addressed all valid criticisms raised, at this point we see little in these criticisms that stands against the applicability of the connectionist approach in principle. Indeed, the arguments raised in these papers do not, in general, reflect a full appreciation of the capabilities of connectionist networks in quasi-regular domains. For example, Pinker (1991) does not acknowledge that connectionist models of both spelling-to-sound (as shown here and in SM89) and of inflectional morphology (Daugherty & Seidenberg, 1992) show the very frequency-by-regularity interaction that he takes as one of the key indicators of the operation of a (frequency insensitive) rule system and a (frequency sensitive) lexical lookup mechanism.

Indeed, there are several aspects of the empirical data in the domain of inflectional morphology that appear at this point to favor an interpretation in terms of a single, connectionist system that is sensitive to both frequency and consistency. We will consider here one such aspect, namely the historical evolution of the English past tense system. Hare and Elman (submitted) have reviewed the pattern of change from the early Old English (EOE) period (*circa* 870) to the present. In EOE, there were two main types of verbs—strong and weak—each consisting of several subtypes. Over the period between 870 and the present, the different types of weak verbs coalesced into a single type: the current "regular" past. Many of the strong verbs "regularized," but several of them persist to this day as the various irregular verbs of modern English. The coalescence of the various types of weak verbs into a single type, the pattern of susceptibility to regularization among the strong verbs, and the occasional occurrence of "irregularization," in which a particular weak verb took on the characteristics of a cluster of strong verbs, are all traced to workings of a single connectionist system that is sensitive both to frequency and consistency. Language change is thought of, in Hare and Elman's approach, as the iterative application of a new generation of learners (simulated by new, untrained networks) to the output of the previous generation of learners (simulated by old networks, trained on the output of even older networks). Each generation imposes its own distortions on the corpus: among these are the elimination of subtle differences between variations of the weak past that apply to similar forms, and the regularization of low-frequency irregular forms with few friends. Gradually over the course of generations, the system is transformed from the highly complex system of *circa* 870 to the much simpler system that is in use today. The remaining irregular verbs are either highly consistent with their neighbors, highly frequent, or both; less frequent and less consistent strong verbs have been absorbed by the regular system. Crucially for our argument, both the "regular" (or weak) system and the "exception" (or strong) system show effects of frequency and consistency, as would be expected on a single-system account.

More generally, we believe that the approach taken here will turn out to be relevant to a wide range of other domains that have a quasi-regular structure, in the sense that they involve systematicity that coexists with some arbitrariness and many exceptions. Derivational morphology presents a rich quasi-regular system. First of all, there are many morphemes that are partially productive in ways that are similar to quasi-regular correspondences in inflectional morphology and spelling-to-sound: that is, they appear to be governed by a set of "soft" constraints. Second, the meaning of a morphologically complex word is related to, but not completely determined by, its constituent morphemes; thus, there is partial, but not complete, regularity in the mapping from meaning to sound (see Bybee, 1985, for a discussion of these points). Even further from the present domain of spelling-to-sound correspondence are the domains encompassed by semantic, episodic, and encyclopedic knowledge. All of these domains are quasi-regular, in that facts and experiences are partially arbitrary, but also partially predictable from the characteristics of other, related facts and experiences (see McClelland, McNaughton, & O'Reilly, 1994, for discussion). Consider the robin, for example. Its properties are largely predictable from the properties of other birds, but its color and exact size, the sound that it makes, the color of its eggs, etc, are relatively arbitrary. Rumelhart (1990; Rumelhart & Todd, 1993) shows how a connectionist network can learn the contents of a semantic network, capturing both the shared structure that is present in the set of concepts—so as to allow generalization to new examples—while at the same time mastering the idiosyncratic properties of particular examples. As another example, consider John F. Kennedy's assassination. There were several arbitrary aspects, such as the date and time of the event, etc. But our understanding of what happened depends on knowledge derived from other events involving presidents, motorcades, rifles, spies, etc. Our understanding of these things informs, indeed pervades, our memory of Kennedy's assassination. And our understanding of other similar events is ultimately influenced by what we learn about Kennedy's assassination. St. John (1992) provides an example of a connectionist network that learns the characteristics of events and applies them to other, similar events, using just the same learning mechanism, governed by the same principles of combined frequency and consistency sensitivity, as our spelling-to-sound simulations.

In summary, quasi-regular systems like that found in the English spelling-to-sound system appear to be pervasive, and there are several initial indications that connectionist networks sensitive to frequency and consistency will provide insight into the way such systems are learned and represented.

Conclusions

At the end of their paper, Coltheart et al. (1993) reach a conclusion that seems to them "inescapable."

> Our ability to deal with linguistic stimuli we have not previously encountered . . . can only be explained by postulating that we have learned systems of general linguistic rules, and our ability at the same time to deal correctly with exceptions to these rules . . . can only be explained by postulating the existence of systems of word-specific lexical representations. (p. 606)

We have formulated a connectionist approach to knowledge and processing in quasi-regular domains, instantiated it in the specific domain of English word reading, and demonstrated that it can account for the basic abilities of skilled readers to handle correctly both regular and exception items while still generalizing well to novel items. Within the approach, the proficiency of humans in quasi-regular domains stems not from the existence of separate rule-based and item-specific mechanisms, but from the fact that the cognitive system adheres to certain general principles of computation in neural-like systems.

Our connectionist approach not only addresses these general reading abilities, but also provides insight into the detailed effects of frequency and consistency both in the naming latency of normal readers, and in the impaired naming accuracy of acquired and developmental dyslexic readers. A mathematical analysis of a simplified system, incorporating only some of the relevant principles, forms the basis for understanding the intimate relationship between these factors and, in particular, the inherently graded nature of spelling-sound consistency.

The more general lexical framework for word reading on which the current work is based contains a semantic pathway in addition to a phonological pathway. In contrast to the lexical and sublexical procedures in dual-route theories, which operate in fundamentally different ways, the two pathways in the current approach operate according to a common set of computational principles. As a result, the nature of processing in the two pathways is intimately related. In particular, a consideration of the pattern of impaired and preserved abilities in acquired surface dyslexia leads to a view in which there is a partial division of labor between the two pathways. The contribution of the phonological pathway is a graded function of frequency and consistent; items weak on both measures are processed particularly poorly. Overt accuracy on these is not compromised, however, because the semantic pathway can pronounce all words (but not nonwords). The relative capabilities of the two pathways is open to individual differences, and these differences may become manifest in the pattern and severity of reading impairments following brain damage.

Needless to say, much remains to be done. The current simulations have specific limitations, such as the restriction to uninflected monosyllables and lack of attention paid to the development of orthographic representations, that need to be remedied in future work. Furthermore, the nature of processing within the semantic pathway has been characterized only in the coarsest way. Finally, a wide range of related empirical issues, including phonological dyslexia, developmental dyslexia, lexical decision, and pseudohomophone and blocking effects, have been addressed only in very general terms. Nonetheless, the results reported here, along with those of others taking similar approaches, clearly suggest that the computational principles of connectionist modeling can lead to a deeper understanding of the central empirical phenomena in word reading in particular, and in quasi-regular domains more generally.

Appendix 1: Accepted Pronunciations of Glushko's (1979) Nonwords

Regular Nonwords		Exception Nonwords	
Nonword	Pronunciation(s)	Nonword	Pronunciation(s)
BEED	/bEd/	BILD	/bIld/, /bild/
BELD	/beld/	BINT	/bInt/, /bint/
BINK	/biNk/	BLEAD	/blEd/, /bled/
BLEAM	/blEm/	BOOD	/bUd/, /b∧d/, /bud/
BORT	/bOrt/	BOST	/bOst/, /b∧st/, /bost/
BROBE	/brOb/	BROVE	/brOv/, /brUv/, /br∧v/
CATH	/k@T/, /kaT/	COSE	/kOs/, /kOz/, /kUz/
COBE	/kOb/	COTH	/kOT/, /koT/
DOLD	/dOld/, /dald/	DERE	/dAr/, /dEr/, /dur/
DOON	/dUn/	DOMB	/dOm/, /dUm/, /dam/, /damb/
DORE	/dOr/	DOOT	/dUt/, /dut/
DREED	/drEd/	DROOD	/drUd/, /dr∧d/, /drud/
FEAL	/fEl/	FEAD	/fEd/, /fed/
GODE	/gOd/	GOME	/gOm/, /g∧m/
GROOL	/grUl/, /grul/	GROOK	/grUk/, /gruk/
HEAN	/hEn/	HAID	/h@d/, /hAd/, /hed/
HEEF	/hEf/	HEAF	/hEf/, /hef/
HODE	/hOd/	HEEN	/hEn/, /hin/
HOIL	/hYl/	HOVE	/hOv/, /hUv/, /h∧v/
LAIL	/lAl/	LOME	/lOm/, /l∧m/
LOLE	/lOl/	LOOL	/lUl/, /lul/
MEAK	/mAk/, /mEk/	MEAR	/mAr/, /mEr/
MOOP	/mUp/	MONE	/mOn/, /m∧n/, /mon/
MUNE	/mUn/, /myUn/	MOOF	/mUf/, /muf/
NUST	/n∧st/	NUSH	/n∧S/, /nuS/
PEET	/pEt/	PILD	/pIld/, /pild/
PILT	/pilt/	PLOVE	/plOv/, /plUv/, /pl∧v/
PLORE	/plOr/	POMB	/pOm/, /pUm/, /pam/, /pamb/
PODE	/pOd/	POOT	/pUt/, /put/
POLD	/pOld/, /pald/	POVE	/pOv/, /pUv/, /p∧v/
PRAIN	/prAn/	PRAID	/pr@d/, /prAd/, /pred/
SHEED	/SEd/	SHEAD	/SEd/, /Sed/
SOAD	/sOd/, /sod/	SOOD	/sUd/, /s∧d/, /sud/
SPEET	/spEt/	SOST	/sOst/, /s∧st/, /sost/
STEET	/stEt/	SPEAT	/spAt/, /spEt/, /spet/
SUFF	/s∧f/	STEAT	/stAt/, /stEt/, /stet/
SUST	/s∧st/	SULL	/s∧l/, /sul/
SWEAL	/swEl/	SWEAK	/swAk/, /swEk/
TAZE	/tAz/	TAVE	/t@v/, /tAv/, /tav/
WEAT	/wAt/, /wEt/, /wet/	WEAD	/wEd/, /wed/
WOSH	/waS/	WONE	/wOn/, /w∧n/, /won/
WOTE	/wOt/	WULL	/w∧l/, /wul/
WUFF	/w∧f/	WUSH	/w∧S/, /wuS/

Appendix 2: Regularizations of Taraban and McClelland's (1987) Exception Words

High-Frequency Exceptions			Low-Frequency Exceptions		
Word	Correct	Regularization(s)	Word	Correct	Regularization(s)
ARE	/ar/	/Ar/	BOWL	/bOl/	/bWl/
BOTH	/bOT/	/boT/	BROAD	/brod/	/brOd/
BREAK	/brAk/	/brEk/	BUSH	/buS/	/b∧S/
CHOOSE	/CUz/	/CUs/	DEAF	/def/	/dEf/
COME	/k∧m/	/kOm/	DOLL	/dal/	/dOl/
DO	/dU/	/dO/, /da/	FLOOD	/fl∧d/	/flUd/, /flud/
DOES	/d∧z/	/dOz/, /dOs/	GROSS	/grOs/	/gros/, /gras/
DONE	/d∧n/	/dOn/	LOSE	/lUz/	/lOs/, /lOz/
FOOT	/fut/	/fUt/	PEAR	/pAr/	/pEr/
GIVE	/giv/	/gIv/	PHASE	/fAz/	/fAs/
GREAT	/grAt/	/grEt/	PINT	/pInt/	/pint/
HAVE	/hav/	/hAv/	PLOW	/plW/	/plO/
MOVE	/mUv/	/mOv/	ROUSE	/rWz/	/rWs/
PULL	/pul/	/p∧l/	SEW	/sO/	/sU/
PUT	/put/	/p∧t/	SHOE	/SU/	/SO/
SAID	/sed/	/sAd/	SPOOK	/spUk/	/spuk/
SAYS	/sez/	/sAz/, /sAs/	SWAMP	/swamp/	/sw@mp/
SHALL	/Sal/	/Sol/	SWARM	/swOrm/	/swarm/
WANT	/want/	/w@nt/	TOUCH	/t∧C/	/tWC/
WATCH	/waC/	/w@C/	WAD	/wad/	/w@d/
WERE	/wur/	/wEr/	WAND	/wand/	/w@nd/
WHAT	/w∧t/	/w@t/	WASH	/woS/	/w@S/
WORD	/wurd/	/wOrd/	WOOL	/wul/	/wUl/
WORK	/wurk/	/wOrk/	WORM	/wurm/	/wOrm/

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451.

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

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.

Baddeley, A. D., Ellis, N. C., Miles, T. C., & Lewis, V. J. (1982). Developmental and acquired dyslexia: A comparison. *Cognition*, *11*, 185–199.

Balota, D. & Ferraro, R. (1993). A dissociation of frequency and regularity effects in pronunciation performance across yong adults, older adults, and individuals with senile dementia of the Alzheimer type. *Journal of Memory and Language*, *32*, 573–592.

Baluch, B. & Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and nonlexical routines in oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 644–652.

Baron, J. & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 207–214.

Beauvois, M.-F. & Derouesné, J. (1979). Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery and Psychiatry*, *42*, 1115–1124.

Behrmann, M. & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, a single lexicon. *Cognitive Neuropsychology*, *9*(3), 209–258.

Besner, D. & Smith, M. C. (1992). Models of visual word recognition: When obscuring the stimulus yields a clearer view. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 468–482.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary? *Psychological Review*, *97*(3), 432–446.

Brousse, O. & Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 380–387.

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

Bub, D. N., Black, S., Hampson, E., & Kertesz, A. (1988). Semantic encoding of pictures and words: Some neuropsychological observations. *Cognitive Neuropsychology*, *5*(1), 27–66.

Buchanan, L. & Besner, D. (1993). Reading aloud: Evidence for the use of a whole word nonsemantic pathway. *Canadian Journal of Experimental Psychology*, *47*(2), 133–152.

Bullinaria, J. A. (submitted). Representation, learning, generalization and damage in neural network models of reading aloud.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia, PA: John Benjamins.

Bybee, J. L. & Slobin, D. L. (1982). Rules and schemas in the development and use of the English past tense. *Language*, *58*, 265–289.

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

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing*. New York: Academic Press.

Coltheart, M. (1985). Cognitive neuropsychology and the study of reading. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3–37.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589–608.

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

Coslett, H. B. (1991). Read but not write "idea": Evidence for a third reading mechanism. *Brain and Language*, *40*, 425–443.

Cottrell, G. W. & Plunkett, K. (1991). Learning the past tense in a recurrent network: Acquiring the mapping from meaning to sounds. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 328–333.

Daugherty, K. & Seidenberg, M. S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 259–264.

Derthick, M. (1990). Mundane reasoning by settling on a plausible model. *Artificial Intelligence*, *46*, 107–157.

Ellis, A. W. & Marshall, J. C. (1978). Semantic errors or statistical flukes? A note on Allport's "On knowing the meanings of words we are unable to report". *Quarterly Journal of Experimental Psychology*, *30*, 569–575.

Fera, P. & Besner, D. (1992). The process of lexical decision: More words about a parallel distributed processing model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 749–764.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Forster, K. (in press). Elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*.

Forster, K. I. & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behaviour*, *12*, 627–635.

Frederiksen, J. R. & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 361–379.

Friedman, R. B. (in press). Phonological dyslexia is a continuum (with deep dyslexia as its endpoint). *Brain and Language*.

Funnell, E. (1983). Phonological processing in reading: New evidence from acquired dyslexia. *British Journal of Psychology*, *74*, 159–180.

Glosser, G. & Friedman, R. B. (1990). The continuum of deep/phonological alexia. *Cortex*, *26*, 343–359.

Gluck, M. A. & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 674–691.

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.

Hanson, S. J. & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, *13*, 471–518.

Hare, M. & Elman, J. L. (1992). A connectionist account of English inflectional morphology: Evidence from language change. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 265–270.

Hare, M. & Elman, J. L. (submitted). Learning and morphological change.

Harris, M. & Coltheart, M. (1986). *Language processing in children and adults*. London: Routledge & Kegan Paul.

Henderson, L. (1982). *Orthography and word recognition in reading*. London: Academic Press.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 161–188.

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

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*(1), 47–76.

Hinton, G. E. (Ed.) (1991). *Connectionist symbol processing*. Cambridge, MA: MIT Press.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press, 77–109.

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

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

Hoeffner, J. (1992). Are rules a thing of the past? The acquisition of verbal morphology by an attractor network. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 861–866.

Hoeffner, J. H. & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In *Proceedings of the 25th Annual Child Language Research Forum*. Stanford, CA, 38–49.

Humphreys, G. W. & Evett, L. J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading. *Behavioral and Brain Sciences*, *8*, 689–740.

Huttenlocher, P. & Huttenlocher, J. (1973). A study of children with hyperlexia. *Neurology*, *26*, 1107–1116.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*, 295–307.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*, 687–715.

Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, *24*, 1–19.

Jordan, M. I. & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354.

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

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathmatical Statistics*, *22*, 79–86.

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

Lacouture, Y. (1989). From mean squared error to reaction time: A connectionist model of word recognition. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceeedings of the 1988 connectionist models summer school*. San Mateo, CA: Morgan Kauffman, 371–378.

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

MacWhinney, B. & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, *40*, 121–153.

Manis, F. R., Seidenberg, M. S., Doi, L. M., McBride-Chang, C., & Peterson, A. (submitted). On the bases of two subtypes of developmental dyslexia.

Marcel, T. (1980). Surface dyslexia and beginning reading: A revised hypothesis of the pronunciation of print and its impairments. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia*. London: Routledge & Kegan Paul, 227–258.

Marchman, V. A. (1993). Constraints on plasticity in a connectionist model of the English past tense. *Journal of Cognitive Neuroscience*, *5*(2), 215–234.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, J. T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–165.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, *202*, 437–470.

Marshall, J. C. (1984). Toward a rational taxonomy of the developmental dyslexias. In R. N. Malatesha & H. A. Whitaker (Eds.), *Dyslexia: A global issue*. The Hague: Martinus Nijhoff, 211–232.

Marshall, J. C. & Newcombe, F. (1966). Syntactic and semantic errors in paralexia. *Neuropsychologia*, *4*, 169–176.

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

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

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*, 398–421.

Masterson, J. (1985). On how we read non-words: Data from different populations. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia*. Hillsdale, NJ: Lawrence Erlbaum Associates, 289–299.

McCann, R. S. & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation and the locus of the word-frequency effects in word naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 14–24.

McCarthy, R. & Warrington, E. K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex*, *22*, 359–380.

McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3–36.

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

McClelland, J. L. (1993). The GRAIN model: A framework for modeling the dynamics of information processing. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience*. Hillsdale, NJ: Lawrence Erlbaum Associates, 655–688.

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. (1994). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory (Technical Report PDP.CNS.94.1). Pittsburgh, PA: Department of Psychology, Carnegie Mellon University.

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., & 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.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, *2*(6), 387–395.

Mehegan, C. C. & Dreifuss, F. E. (1972). Hyperlexia: Exceptional reading in brain-damaged children. *Neurology*, *22*, 1105–1111.

Metsala, J. L. & Siegel, L. S. (1992). Patterns of atypical reading development: Attributes and underlying reading processes. In S. J. Segalowitz & I. Rapin (Eds.), *Handbook of neuropsychology*, Vol. 7. Amsterdam: Elsevier Science Publishers.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1974). Functions of graphemic and phonemic codes in visual word recognition. *Memory and Cognition*, *2*, 309–321.

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

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 148–197.

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 452–467.

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, *24*, 45–64.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, *7*, 323–331.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Morton, J. & Patterson, K. (1980). A new attempt at an interpretation, Or, an attempt at a new interpretation. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia*. London: Routledge & Kegan Paul, 91–118.

Movellan, J. R. & McClelland, J. L. (1991). Learning continuous probability distributions with the contrastive hebbian algorithm (Technical Report PDP.CNS.91.2). Pittsburgh, PA: Department of Psychology, Carnegie Mellon Univeirsity.

Mozer, M. C. (1990). Discovering faithful "Wickelfeature" representations in a connectionist network. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Norris, D. (submitted). A quantitative multiple-levels model of reading aloud.

Olsen, A. & Caramazza, A. (1991). The role of cognitive theory in neuropsychological research. In S. Corkin, J. Grafman, & F. Boller (Eds.), *Handbook of neuropsychology*. Amsterdam: Elsevier, 287–309.

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

Patterson, K., Graham, N., & Hodges, J. R. (1994). Reading in Alzheimer's type dementia: A preserved ability? *Neuropsychology*, *8*(3), 395–412.

Patterson, K. & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia*, *30*(12), 1025–1040.

Patterson, K. E. (1990). Alexia and neural nets. *Japanese Journal of Neuropsychology*, *6*, 90–99.

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

Patterson, K. E. & Marcel, A. J. (1992). Phonological ALEXIA or PHONOLOGICAL alexia? In J. Alegria, D. Holender, J. Junça de Morais, & M. Radeau (Eds.), *Analytic approaches to human cognition*. New York: Elsevier, 259–274.

Patterson, K. E. & Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia*. Hillsdale, NJ: Lawrence Erlbaum Associates, 335–359.

Patterson, K. E., Seidenberg, M. S., & McClelland, J. L. (1990). 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*. London: Oxford University Press, 131–181.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*(2), 263–269.

Phillips, W. A., Hay, I. M., & Smith, L. S. (1993). Lexicality and pronunciation in a simulated neural net (Technical Report CCCN-14). Stirling, Scotland: Centre for Cognitive and Computational Neuroscience, University of Stirling.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

Pinker, S. & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D. C. (in press). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*.

Plaut, D. C. & Hinton, G. E. (1987). Learning sets of filters using back propagation. *Computer Speech and Language*, *2*, 35–61.

Plaut, D. C. & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 824–829.

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

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

Plunkett, K. & Marchman, V. A. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*(1), 21–69.

Pugh, K. R., Rexer, K., & Katz, L. (1994). Evidence of flexible coding in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 807–825.

Quinlan, P. (1991). *Connectionism and psychology: A psychological perspective on new connectionist research*. Chicago: University of Chicago Press.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quartely*, *27*, 29–53.

Reggia, J. A., Marsland, P. M., & Berndt, R. S. (1988). Competitive dynamics in a dual-route connectionist model of print-to-sound transformation. *Complex Systems*, *2*, 509–547.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks*. New York: Academic Press, Chap. 21.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (in press). Backpropagation: The basic theory. In D. E. Rumelhart & Y. Chauvin (Eds.), *Backpropagation: Theory and practice*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press, 318–362.

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

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

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*. Cambridge, MA: MIT Press, 216–271.

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

Rumelhart, D. E. & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience*. Cambridge, MA: MIT Press, 3–30.

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

Schwartz, M. F., Marin, O. M., & Saffran, E. M. (1979). Dissociations of language function in dementia: A case study. *Brain and Language*, *7*, 277–306.

Schwartz, M. F., Saffran, E. M., & Marin, O. S. M. (1980). Fractioning the reading process in dementia: Evidence for word-specific print-to-sound associations. In M. Coltheart, K. E. Patterson, & J. C. Marshall (Eds.), *Deep dyslexia*. London: Routledge & Kegan Paul, 259–269.

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

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

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

Seidenberg, M. S. & McClelland, J. L. (1990). More words but still no lexicon: Reply to Besner et al. (1990). *Psychological Review*, *97*(3), 477–452.

Seidenberg, M. S. & McClelland, J. L. (1992). Connectionist models and explanatory theories in cognition (Technical Report PDP.CNS.92.4). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

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

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behaviour*, *23*, 383–404.

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

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shallice, T. & McCarthy, R. (1985). Phonological reading: From patterns of impairment to possible procedures. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia*. Hillsdale, NJ: Lawrence Erlbaum Associates, 361–398.

Shallice, T. & McGill, J. (1978). The origins of mixed errors. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Lawrence Erlbaum Associates, 193–208.

Shallice, T., Warrington, E. K., & McCarthy, R. (1983). Reading without semantics. *Quarterly Journal of Experimental Psychology*, *35A*, 111–138.

Silverberg, N. E. & Silverberg, M. C. (1967). Hyperlexia—Specific word recognition skills in young children. *Exceptional Children*, *34*, 41–42.

St. John, M. F. (1992). The story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*, 271–306.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press, 444–459.

Strain, E., Patterson, K. E., & Seidenberg, M. S. (1994). Semantic effects in single word naming (Technical Report PDP.CNS.94.3). Pittsburgh, PA: Department of Psychology, Carnegie Mellon University.

Sutton, R. S. (1992). Adapting bias by gradient descent: An incremental version of Delta-Bar-Delta. In *Proceedings of the 11th National Conference on Artificial Intelligence*.

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

Treiman, R. & Chafetz, J. (1987). Are there onset- and rime-like units in printed words? In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, 281–327.

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

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

Vellutino, F. (1979). *Dyslexia*. Cambridge, MA: MIT Press.

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

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

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

Widrow, G. & Hoff, M. E. (1960). Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4*, 96–104.

Williams, R. J. & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, *2*(4), 490–501.

Zorzi, M., Houghton, G., & Butterworth, B. (1994, January). Phonological reading: A connectionist investigation. Paper Presented at the Twelfth European Workshop on Cognitive Neuropsychology, Bressanone, Italy.

Zorzi, M., Houghton, G., & Butterworth, B. (in preparation). Two routes or one in reading aloud? A connectionist "dual-process" model.