ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cognitive Psychology 45 (2002) 413–445

Cognitive
Psychology

www.academicpress.com

Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand

Jeffrey S. Bowers

Department of Experimental Psychology, University of Bristol, Bristol BS8-1TN, UK

Accepted 29 April 2002

Abstract

One of the central claims associated with the parallel distributed processing approach popularized by D.E. Rumelhart, J.L. McClelland and the PDP Research Group is that knowledge is coded in a distributed fashion. Localist representations within this perspective are widely rejected. It is important to note, however, that connectionist networks can learn localist representations and many connectionist models depend on localist coding for their functioning. Accordingly, a commitment to distributed representations should be considered a specific theoretical claim regarding the structure of knowledge rather than a core principle, as often assumed. In this paper, it is argued that there are fundamental computational and empirical challenges that have not yet been addressed by distributed connectionist theories that are readily accommodated within localist approaches. This is highlighted in the context of modeling word and nonword naming, the domain in which some of the strongest claims have been made. It is shown that current PDP models provide a poor account of naming monosyllable items, and that distributed representations make it difficult for these models to scale up to more complex language phenomena. At the same time, models that learn localist representations are shown to hold promise in supporting many of the core reading and language functions on which PDP models fail. It is concluded that the common rejection of localist coding schemes within connectionist architectures is premature.

E-mail address: j.bowers@bris.ac.uk.

0010-0285/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S0010-0285(02)00506-6

1. Introduction

Since the publication of the two volume set *Parallel Distributed Processing* (McClelland, Rumelhart, & the PDP Research Group, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986), connectionist models have played a central role in theorizing about perception, memory, language, and cognition more generally. This approach has reintroduced learning as a core constraint in theory development, it shows promise of identifying a set of general principles that apply across a wide range of cognitive domains, and it provides a possible bridge between theories of cognition and the neural structures that mediate these functions. And by linking theories of cognition with theories of learning and neurobiology, connectionism holds the promise of identifying principled constraints as to why cognitive systems are organized the way they are as opposed to other plausible alternatives; that is, this approach shows promise of supporting *explanatory* rather than *descriptive* theories (Seidenberg, 1993a).

Despite these general appeals, two central claims associated with the PDP agenda remain controversial. The first—and the focus of the present paper—is the claim that knowledge is coded in a distributed fashion. That is, words, objects, simple concepts (e.g., DOG), etc. are assumed to be coded as a pattern of activation across many processing units, with each unit contributing to many different representations, or as Hinton, McClelland, and Rumelhart (1986) put it: "Each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" (p. 77, Vol. 1). This contrasts with so-called "localist" representations, in which each unit represents something meaningful (e.g., a semantic node for *grandmother*), with distinct units encoding for distinct pieces of information (e.g., a separate node for *grandfather*). The rejection of localist coding schemes in favour of distributed representations is sometimes described as one of the "core connectionist principles" (e.g., Seidenberg, 1993b, p. 300), or "general connectionist principles" (e.g., Plaut & Shallice, 1993, p. 377). Introductory textbooks often make similar claims:

> Traditional models of cognitive processing usually assume a local representation of knowledge. That is, knowledge about different things is stored in different, independent locations... In connectionist models information storage is not local, it is distributed. There is no one place where a particular piece of knowledge can be located" (McLeod, Plunkett, & Rolls, 1998, p. 31).

At this point, the association between distributed coding schemes and connectionism is so strong that little consideration is given to connectionist models that learn local representations—to the point that authors often provide definitions of connectionism that exclude localist connectionist models that learn (as in the definitions quoted above).

The second contentious claim is that human cognition is achieved without recourse to explicit rules, a view that is sometimes referred to as "eliminative connectionism" (Pinker & Prince, 1988). That is, it is assumed that the mind is a statistical learning device that encodes the structure of the environment in such a way that it supports "rule-like" behavior, but not rules per se. (e.g., McClelland & Plaut, 1999). Central to implementing a rule system is the idea that abstract and context independent categories exist (e.g., *noun*, *dog*, or *Fido*), and that computations are performed over these representations. This contrasts with the PDP approach that assumes that computations are performed over individual tokens that are directly presented to the network. On this view internal representations can be learned (whenever hidden units are included) but they are neither abstract nor context independent, as outlined in more detail below.

Together, the rejection of localist representations and rule systems challenges two of the most fundamental assumptions of more traditional "symbolic" approaches to cognitive theorizing. Given these strong claims, it is not surprising that there has been an active and ongoing debate concerning these issues since the publication of the PDP books. The most active discussion has focused on the role of rules (if any) in cognition (with Fodor, 2000; Fodor & Pylyshyn, 1988; Marcus, 2001; Pinker, 1999, among others arguing yes to rules; and McClelland & Plaut, 1999; McClelland & Seidenberg, 2000, amongst others arguing no), with no signs of letup. The corresponding debate concerning the relative merits of localist and distributed coding schemes has attracted somewhat less attention, although the arguments have been no less heated (e.g., Page, 2000; and accompanying responses). The relative lack of attention to this latter issue may be attributed, in part, to the fact that the local-distributed debate has largely been considered within a rather specialized context: namely, in developing theories that support the identification and naming of monosyllable word and nonwords from print (for advocates of localist coding, see Besner, 1999; Besner, Twilley, McCann, & Seergobin, 1990; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger & Jacobs, 1998; for advocates of distributed coding, see Harm & Seidenberg, 1999; Plaut & McClelland, 2000; Seidenberg & McClelland, 1990). Nevertheless, this local-distributed contrast has general implications concerning the structure of knowledge in the mind (brain) with associated consequences for developing models in various domains. Indeed, as discussed below, the rules/non-rules and localist/distributed contrasts are not unrelated, and connectionist networks that implement rules (e.g., Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Shastri & Ajjanagadde, 1993) rely on localist coding schemes.

The present paper focuses on the first debate, and challenges the widespread assumption that distributed coding schemes enjoy advantages over localist coding schemes within connectionist systems. Page (2000) has recently provided a strong argument in support of localist coding schemes, focusing on the strengths of connectionist models that learn localist representation. Among other things, he demonstrates that these models can support a wide range of phenomena based on their close relationship to a number of classical mathematical models of behavior, and argues that standard criticisms of localist models—i.e., that they do not generalize, do not degrade gracefully, are not biologically plausible, etc.—are all unfounded. In the present paper, I take the complementary tack, and focus on the weakness of distributed coding schemes. In particular, past evidence taken in support of distributed coding schemes is challenged, the functional utility of these codes is questioned, and computational limitations of all current models that reject localist representations are identified. These challenges are highlighted within the context of theories of language, in most cases reading. At the same time, I contrast the familiar PDP approach with connectionist models that learn localist representations in order to make the reader aware of this alternative framework, as well as to demonstrate its promise. Hopefully, along with Page (2000), this critique will help generate a debate concerning the relative merits of connectionist models that learn distributed and localist representations, and encourage researchers to consider a wider range of connectionist models—such as the models developed and inspired by Grossberg and his colleagues.

2. Background issues

Before raising any specific arguments, I would like to set the stage by identifying some possible points of confusion. First, it is important to be clear what constitutes a connectionist network. Although this might seem self-evident, the issue is complicated by the role that learning plays in defining the term. Spreading activation theories of semantic memory (e.g., Collins & Loftus, 1975) as well as the logogen and the interactive activation models of word identification (McClelland & Rumelhart, 1981; Morton, 1979) should be considered connectionist models with localist coding schemes if the term refers to any model that represents information as a pattern of activation across a set of interconnected units. But if a central and defining feature of connectionist models is that they support learning, then these and more recent hand-wired localist models by Dell (1986), Grainger and Jacobs (1996) and others can be excluded from this category.

But even if one agrees that learning is one of the defining features of connectionist models, a critical point that needs to be emphasized is that connectionist models can learn localist representations. This, however, is not widely acknowledged. For example, in a response to an paper by Page (2000) in which localist coding schemes were advocated, Plaut and McClelland (2000) wrote:

> …whereas Page would stipulate localist representations for various types of problems, our approach allows an appropriate representation to be created in response to the constraints built into the learning procedure and the task at hand. (p. 490)

Similar points have been published elsewhere. More commonly, PDP models that learn distributed representations are contrasted with localist models that are hand-wired.

However, this characterization of localist networks is mistaken. For example, Grossberg and colleagues have developed networks that learn localist representations without any need to define categories a priori (e.g., Carpenter & Grossberg, 1987; Grossberg, 1980). Indeed, Grossberg (1987) specifically criticized the McClelland and Rumelhart (1981) model that did include hand-wired "letter" and "word" units, and argued that it is preferable to employ the more abstract terms "item" and "list" to refer to learned units that are not pre-specified. A *list* level simply codes for a collections of items. So for example, nodes within the list level will learn to represent commonly occurring groupings of items, which may include words, affixes (e.g., -*ed*, -*ing*, *un*-), stem morphemes (e.g., *ject*, *vise*), bodies (e.g., *ead*, *ind*), and even letters. Grossberg's models were discussed by Page (2000) and have been extended (as well as implemented) by Nigrin (1993) and, with specific focus on visual word identification, by Davis (1999). It is not true that localist representations need to be hand-wired on some pre-theoretical basis.

The implication is important in the present context: Distributed representations are not an intrinsic property of connectionist systems that learn. Instead, it is a specific theoretical claim regarding the structure of knowledge within this more general framework, and it is reasonable to ask under what conditions high-level entities (e.g., words) are represented in a distributed format (if ever), and under what conditions localist representations develop (if ever).

A second general issue of possible confusion is with regard to the types of localist representations that are rejected on the PDP approach. The defining feature of a localist representation is simple enough: Localist coding schemes include separate representations (nodes in a connectionist network) for distinct pieces of information. So for example, when a single node in a network codes for the written letter *A* and a second node codes for the letter B, the network has encoded these letters in a localist format.[1] The possible confusion relates to the level at which knowledge—orthographic or other-

[1] Localist coding would also be implemented if the letter A was coded with a collection of nodes that did not overlap with nodes involved in representing other letters. So, in terms of implementing a localist coding scheme in neural hardware, one is not committed to assuming that a single neuron codes for a complex piece of information. But one is committed to the view that there is some collection of neurons uniquely involved in coding for the letter A, and another set of non-overlapping neurons uniquely involved in coding for B, etc.

wise—is claimed to be distributed within the PDP framework. Indeed, within this approach, the extent to which models include distributed representations varies considerably. For example, on some models of word identification, orthographic and phonological knowledge are coded in a distributed format at all levels, from the input to output units, and there is no single unit in the model that uniquely defines any piece of meaningful information (Seidenberg & McClelland, 1989). On other models, individual phonetic features, phonemes, letters, or even complex graphemes (e.g., the letter string *TCH*) are coded in a localist format (e.g., Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg, & Patterson, 1996). But in all cases, knowledge at the lexical level is coded in a distributed format, and this is the key theoretical claim that many authors make, as can be seen in the following quote:

> The present model departs from these precursors in a fundamental way: Lexical memory does not consist of entries for individual words; there are no logogens. Knowledge of words is embedded in a set of weights on the connections between processing units encoding orthographic, phonological, and semantic properties of words, and the correlations between these properties.... Thus, the notion of lexical access does not play a central role in our model because it is not congruent with the model's representational and processing assumptions (Seidenberg & McClelland, 1989, p. 560).

As noted by Page (2000; also see Zorzi, Houghton, & Butterworth, 1998), the willingness to include localist sub-lexical codes within a network while rejecting localist lexical representations is surprising, as it leads to the situation in which complex graphemes such as TCH are represented locally whereas high-frequency words such as THE are not (Plaut et al., 1996). But in any case, the central claim of the present paper is that the rejection of localist representations at the lexical level leads to serious limitations that have not yet been confronted.

The above considerations should also highlight the fact that "stipulation" is not restricted to models with localist coding schemes. Theorists committed to distributed lexical knowledge sometimes stipulate localist grapheme and phoneme units (Plaut et al., 1996), or localist letter and phonetic feature units (Harm & Seidenberg, 1999) in order to improve the performance of their models. Similarly, modelers must stipulate the learning algorithm employed: Models that rely on back-propagation learn distributed lexical representations, whereas adaptive resonance (e.g., Carpenter & Grossberg, 1987) and various competitive schemes (e.g., Grossberg, 1976; Rumelhart & Zipser, 1985) learn localist lexical codes. Indeed, given that learning algorithms that support the development of distributed knowledge at the lexical level (e.g., back-propagation) cannot also support the development of localist knowledge at the letter or grapheme levels, qualitatively different learning principles need to be stipulated within the PDP approach. By contrast, the same learning algorithm can support the development of localist letter and lexical representations (e.g., Davis, 1999; Grossberg, 1987), suggesting that the distributed approach involves more not less stipulation. At the very least, the distributed approach includes every bit as much stipulation as the localist approach.

One final point of general introduction should be noted. One of the compelling features of connectionist models that learn distributed representations is that the same set of principles can be used to explain to a wide range of phenomena. To take the example of reading, the same principles have been used to explain the reading of regular and irregular monosyllabic words (Seidenberg & McClelland, 1989) and nonwords (Plaut et al., 1996), as well as various developmental and acquired disorders of reading (e.g., Harm & Seidenberg, 1999; Plaut et al., 1996), semantic priming phenomena (Plaut & Booth, 2000), among other findings. By contrast, more traditional models that include localist representations often rely on qualitatively different mechanisms to solve different tasks. For example, according to the dual route model of reading, irregular words are identified in a network that includes localist lexical representations (such as the Interactive Activation model of McClelland & Rumelhart, 1981) whereas nonwords are read by a set of grapheme-phoneme conversion rules—two systems that operate according to qualitatively different principles (e.g., Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart et al., 2001). The fact that connectionist models with distributed representations can accommodate a wide variety of phenomena is often thought to provide evidence in support of this general approach. As Plaut (1999) puts it: "...their relative success at reproducing key patterns of data in the domain of word reading, and the fact that the very same computational principles are being applied successfully across a wide range of linguistic and cognitive domains, suggests that these models capture important aspects of representation and processing in the human language and cognitive domains" (pp. 362–363).

And indeed, the identification of general principles that apply to a wide range of phenomena is one of the most important functions a theory can fulfill. But for present purposes, the relevant point is that this capacity is not restricted to connectionist models that learn distributed representations. For example, Grossberg and colleagues have developed models of classical and operant conditioning, early vision, visual object recognition, visual word identification, low-level phonology, speech recognition, eye-movement control, working memory, episodic memory, attention shifting, among other phenomena, all employing a small set of principles incorporated within ART and related networks that can learn localist representations (cf. Grossberg, 1999). Others have developed and extended these networks so that models are better suited to segment and learn familiar patterns embedded in larger patterns, as is necessary in speech segmentation (Nigrin, 1993), visual word identification (Davis, 1999), music perception (Page, 1994), among other areas. This is not to say that these models are correct, but rather, that connectionist models with distributed representations should not be preferred because of their success in a variety of domains.

The point of this extended introduction is to demonstrate that there are no general overarching reasons to prefer distributed compared to localist coding schemes: Connectionist models can learn either localist or distributed representations, modelers must "stipulate" various features of their networks, but this applies to models that learn distributed as well as localist representations, and both approaches have been applied to a wide range of phenomena. Furthermore, as shown by Page (2000) the standard criticisms levied against localist coding schemes are ill founded. Accordingly, if one is going to argue that distributed coding schemes are to be preferred over localist representations it is not sufficient to appeal to any of these general considerations. Rather, it is necessary to demonstrate an advantage of distributed coding schemes when the two approaches are directly compared in terms of their capacity to accommodate existing data as well as their promise in scaling up to more realistic settings. Unfortunately, this analysis has largely been lacking from the literature, in part because PDP models of word identification have focused on a restricted set of phenomena (e.g., naming monosyllable items), and in part because it is not generally recognized that connectionist models can learn localist representations in the first place.

The present paper contributes to this analysis by identifying a number of key limitations of current PDP approaches that prevent these models from addressing more complex language phenomena—in particular, the capacity to identify complex word forms and code for multiple items at the same time. These limitations, however, are overcome in connectionist models that learn localist representations, suggesting that this latter approach deserves more serious consideration. But before focusing on these issues, the paper reviews more familiar territory; namely, the success of PDP models in naming and identifying monosyllable words.

3. Identifying and naming monosyllabic words and nonwords

As noted above, the relative merits of localist vs. distributed coding schemes have largely been considered in relation to developing models of naming monsyllable words and nonwords, and it is within this domain that distributed models have enjoyed some notable successes. For example, these models can name nonwords (e.g., *blap*) and irregular words (e.g., *pint*) with a single set of processes. Prior to the work of Seidenberg and McClelland (1989) and Plaut et al. (1996), it was widely assumed that qualitatively different mechanisms were necessary in order to accomplish these two functions. And these same models can accommodate dozens of specific empirical results concerning the processing of single syllable words, based on the same general principles that have been applied to various domains of knowledge.

Still, a number of key phenomena have yet to be explained within this domain (Besner, 1999, lists 10 such findings). Briefly, let me mention two. One limitation not mentioned by Besner is that the current models have not provided an explicit account of the word superiority effect (WSE) in which words are better identified than pseudowords or single letters—classic findings that are often thought to provide strong empirical evidence in support of localist coding within the orthographic system. It is important to note that localist representations support the WSE in the *Interactive Activation Model* of word identification (McClelland & Rumelhart, 1981), as well as more recent versions of this model (Grainger & Jacobs, 1996). Thus it is odd that theories that reject localist coding schemes have not been systematically tested in their ability to accommodate this rich data set. By contrast, connectionist models that learn localist representations do support the WSE (Murre, Phaf, & Wolters, 1992).

But perhaps the most striking problem is that these distributed models account for little variance in reading response times at the item level, despite their capacity to produce the phonology of words and nonwords from print. For example, Spieler and Balota (1997) asked participants to read all the words in the training corpora of the Seidenberg and McClelland (1989) and Plaut et al. (1996) models, and then carried out regression analyses comparing the mean naming latencies of the participants to the performance of the model. When the estimated naming latency of the Seidenberg and McClelland model was compared to the human data, the model accounted for 10.1% of the variance. By contrast, when log frequency, neighborhood density (Coltheart's N), and word length were entered to a regression they accounted for 21.7% of the variance, showing that there was much room for improvement. Further, when the Plaut et al. (1996) model was tested, it only accounted for 3.3% of the variance in word naming latency. So, although the model's performance was improved with regards to its ability to pronounce nonwords, it was at the expense of its ability to predict word naming response latencies at the item level (see Seidenberg & Plaut, 1998, for response). More recently, Coltheart et al. (2001) tested the Plaut et al. (1996) model on a set nonwords included in a study by Weekes (1997), and found that it accounted for less than 1% of the naming latency variance, whereas the Coltheart et al. (1993) dual-route model explained 39% of the RT variance.[2] Thus, although these PDP models can name single syllable words and nonwords, there is little evidence that they accomplish this in the way humans do.

[2] The relative success of the Coltheart et al. model is due to the fact that it predicts increasing naming RTs with increased nonword length, which was observed by Weekes (1997). By contrast, PDP models expect little or no effect of length.

A more general issue merits consideration as well. Although connectionist networks with distributed representations are capable of supporting some forms of generalization—for example, pronouncing single syllable nonwords—the ability to generalize is useless (indeed, undesirable) when arbitrary mappings must be learned. Learning that the word CHAIR refers to a piece of furniture does not provide any information about the meaning of the orthographically related nonword CHAID, and it would be a mistake to generalize in these cases. Thus, one of the advantages of distributed coding schemes, namely, their ability to generalize, is not an asset in these conditions. This seems to undermine much of the adaptive value of distributed coding in the "triangle" model of reading advocated by Seidenberg and colleagues (e.g., Seidenberg & McClelland, 1989) that includes mappings between orthographic-semantic, phonological-semantic, as well as orthographic-phonological representations, with only the latter mappings systematic in a way that can exploit the distributed coding schemes (cf. Forster, 1994).

Note, it is not the case that distributed coding schemes are a neutral format in which to learn arbitrary input-output mappings, but they are actually maladaptive, making learning more difficult as incorrect generalizations tend to be produced (e.g., activating the semantics of CHAIR when the input is CHAID). What improves performance in these situations is the inclusion of more hidden units (e.g., McRae, deSa, & Seidenberg, 1997; Plaut, 1997). For example, in an attempt to model lexical decisions based on the activation of distributed representations within semantics (lexical decision performance is poor if based on the activation of distributed orthographic or phonological codes), Plaut (1997, p. 788) notes: "A much larger number of hidden units was needed to map to semantics than to map from orthography to phonology because there is no systematicity between the surface forms of words and their meaning, and connection networks find unsystematic mappings particularly difficult to learn." Indeed, without including many more hidden units, the model could not learn these mappings. It gets easier when more hidden units are included—and in a model like ARTMAP that has a separate localist units for every word, learning arbitrary mappings is easier still (e.g., Carpenter, Grossberg, & Reynolds, 1991).

Still, despite these difficulties and the lack of any obvious functional utility of distributed coding schemes when mapping between arbitrary domains, it is nevertheless argued that distributed representations mediate these mappings. As far as I am aware, however, the only evidence put forward in support of this claim was reported by Hinton and Shallice (1991) and later by Plaut and Shallice (1993). These authors provided a detailed account of the various reading errors associated with deep dyslexia using a connectionist model that mapped between arbitrarily related orthographic and semantic representations using a distributed coding scheme. After learning these associations, a single lesion to the network caused the model to make a pattern of errors similar to these patients; that is, it made many semantic errors (e.g., settling into the semantic pattern for CAT given the orthographic input DOG), visual errors (e.g., settling into LOG given the input DOG) as well as mixed visual-and-semantic errors (e.g., settling into HOG given DOG) at a rate greater than chance. This pattern of errors was found to be quite general, extending to various network architectures with distributed representations that included recurrent connections. The ability to accommodate this complex pattern of results with a single lesion was an improvement over previous accounts that needed to assume multiple lesion sites. This success was taken to support the conclusion that distributed coding schemes support orthographic-semantic mappings: "We identify four properties of networks that underlie their ability to reproduce the deep dyslexic symptom-complex: *distributed orthographic and semantic representations* (italics added), gradient descent learning, attractors for word meaning, and greater richness of concrete vs. abstract semantics." (Plaut & Shallice, 1993, p. 377).

However, in contrast with the authors initial claim, these results do not depend upon distributed representations. The problem with this claim is that similar patterns of errors are found in connectionist models of speech production (in which information travels from semantics to phonology rather from orthography to semantics) that incorporate localist representations and feedback connections (e.g., Dell, 1986; Dell & O'Seaghdha, 1991). For example, when converting a semantic to a phonological pattern, these networks would on occasion make various forms of phonological errors (e.g., output SHEEP rather than SHEET), various sorts of semantic errors (e.g., CAT rather than DOG), and importantly mixed errors (e.g., RAT rather than CAT) more often than chance. What turns out to be critical in producing these errors is the interactive nature of processing in the network, with information traveling in both the semantic-phonological and phonological-semantic directions. Distributed representations are not relevant.

To summarize, current PDP models of word processing have been designed to support the identification and naming of monosyllabic words and nonwords, but within this domain they face serious empirical limitations. The adaptive value of distributed coding is unclear when mapping between orthographic-semantic or phonological-semantic representations—approximately two thirds of all the mappings within the triangle model proposed by Seidenberg and McClelland (1989). Furthermore, there is no evidence that the distributed codes support these mappings, despite initial claims. Accordingly, the current successes of PDP models do not warrant a strong commitment to distributed representations.

But even if subsequent PDP models overcome these problems, the conclusion that distributed representations underlie reading (and cognition more generally) would be premature. In order to justify these strong claims it is also necessary to demonstrate that this framework shows greater promise than the localist approach in supporting more complex language functions—such as identifying morphologically complex words, or encoding two words at the same time. These latter skills have rarely been considered within PDP literature, but it is within this domain that distributed coding approaches seem most problematic—and where localist coding schemes show the most promise.

4. Identifying complex word forms

One of the important limitations of current PDP models of reading is that they are restricted to monosyllabic words. Given the claim that lexical orthographic and phonological knowledge is coded in a distributed manner, it will be important to demonstrate that this approach can be extended to the processing of more complex word forms, such as novel compound words (e.g., CATPOLE).

However, not only are the input and output coding schemes employed by current PDP models restricted to naming of monosyllable items, there are reasons to think that these approaches will have difficulties scaling up to complex forms. Models that learn distributed lexical codes have used either relational coding schemes in which each letter is coded within a local context, ignoring the absolute position of each letter in a word (Seidenberg & McClelland, 1989) or slot based approaches in which letters or graphemes are explicitly coded in long-term memory in terms of their location within a word (e.g., Harm & Seidenberg, 1999; Plaut et al., 1996). For example, the Wickelcoding scheme used by Seidenberg and McClelland (1989) relies on relational coding in which each letter is coded relative to its immediate context, so that the word TEST would be coded as the Wickelfeatures #TE, TES, EST, ST#, where # refers to a word boundary. The word TEST would be identified when all the relevant Wicklefeatures were identified, without any need to encode the position of these features. In the slot based approach, letter units are tagged in terms of their position within the word. For example, in the Harm and Seidenberg (1999) model, each letter is coded in terms of the position of the letter relative to the vowel, with the vowel coded in position 4 (such that TEST is coded as T3, E4, S5, and T6).

One problem with both relational and slot based coding schemes, however, is that they obscure letter-phoneme correspondences. So for instance, the mapping between T → /t/ is relatively constant across letter positions within a word, but this regularity is lost in the Wickelcoding scheme as the two Ts in TEST are represented by the unrelated orthographic forms #TE and ST#, and this is also true of slot-based approaches, as the two Ts are represented by the units *T-in-third-position* and *T-in-sixth-position*. That is, the regularities are dispersed across unrelated orthographic forms. Plaut et al. (1996) attributed the poor nonword naming performance of the Seidenberg and McClelland (1989) model to the extreme version of the dispersion problem associated with the Wickelcoding scheme (the Ts in TASK and TRIP are coded by the unrelated units #TA and #TR, despite the Ts occurring the in same position). The slot-based coding schemes employed by Plaut et al. (1996) and Harm and Seidenberg (1999) reduced this dispersion (the two Ts are coded by the same unit) allowing these models to pronounce single syllable nonwords.

Still, these latter two models treat the two Ts in the word TEST as unrelated letters, and as such, the learning of T in one context does not apply to T in the other. Although the dispersion problem was reduced to the point that the models could learn to read monosyllable words and nonwords, the problem persists and manifests itself in other ways. One problem noted by Share (1995) is that these models need thousands of learning trials in which an explicit teacher provides correct feedback on every trial—despite the fact that explicit feedback is the exception rather than the rule when a child learns to read; also see (Coltheart et al., 1996). But dispersion becomes more problematic when complex word forms are considered. Consider an example taken from Davis (1999) in which a model has learned the words CAT and POLE, and then is tested on the novel and morphologically complex word CATPOLE. On a slot-based coding scheme in which letters are coded by absolute position, the letters P-O-L-E in CATPOLE are coded as P4, O5, L6, and E7. Accordingly, even though the model has been trained on the monomorphemic word POLE, this training is irrelevant to naming CATPOLE as POLE is coded as P1, O2, L3, end E4—that is, POLE within CATPOLE is orthographically unrelated to POLE by itself. (All slot coding schemes suffer the same problem.) Perhaps the pronunciation of CATPOLE could be supported by learning other words that have letters in the corresponding positions of CATPOLE, such as DIAPER, CHOCOLATE, FUNNEL, and PASSAGE, which have the letters P, O, L, E in positions four, five, six, and seven, respectively. But in any case, this would not help the reader to identify the meaning of CATPOLE, as these latter correspondences are irrelevant to processing the meaning of POLE coded in positions four-to-eight.

One tempting way to address this problem would be to develop a procedure in which only CAT is input initially (i.e., inputting C1, A2, and T3 without P4, O5, L6, and E7) and then only POLE (i.e., treating P1, O2, L3, and E4 as input, excluding C1, A2, and T3). This is similar to a proposal of Plaut (1999) who outlined a model that refixated on different parts of words (i.e., subsets of words would be input to the model, as above) in an attempt to explain the increased naming latencies associated with longer words. But even if this scheme was applied and proved successful in identifying the constituents of novel compound words (CAT and POLE), the solution does not work in general. In order to provide a possible interpretation of a compound (e.g., CATPOLE—a pole for a cat) it is necessary to coactivate the meanings of CAT and POLE (relating POLE to CAT). That is, it is not enough to simply activate the concept CAT one moment, and then replace this with the activation of concept POLE: What would POLE be related to? But the requirement to co-activate both constituents simultaneously is difficult to achieve with distributed representations. These coding schemes only support the activation of one thing at a time, as discussed below.

Note, I am not claiming that distributed representations cannot support the identification of novel compound words in principle. But no one in the PDP community has attempted to address this problem, nor have any potential solutions been suggested. At the same time, networks that learn localist letter and word codes have already been developed that support the identification of complex word structures.

4.1. A possible solution involving a localist coding scheme

The solution depends upon a spatial coding scheme in which localist letter units are coded in long-term memory in a position invariant fashion, with letter position encoded by the temporary pattern of activation across the set of letter units. In particular, sequential letters are coded with decreasing activation values, with a constant activation ratio between successive items, what Grossberg (1978) calls the *invariance principle*. For instance, CATPOLE would be represented as in Fig. 1. One of the consequences of this scheme is that that the same letter nodes are used for each letter, regardless of position. And furthermore, the invariance principle insures that the pattern of activation over the letters P-O-L-E is the same regardless of context (that is, by itself or embedded in CATPOLE), allowing direct access to both CAT and POLE. Note, repeated letters can be encoded within this scheme (see Bradski, Carpenter, & Grossberg, 1994; Davis, 1999). And because CAT and POLE are represented as localist representations at the lexical level, these items can become co-activated the first time they are presented (with CAT more active than POLE), and if these items consistently co-occur, a single lexical representation can be learned.

A number of related constraints can also be satisfied within this framework. For example, models employing spatial coding schemes can correctly identify subset and superset patterns, such that a network can correctly identify the words SELF, the superset MYSELF, and the subset ELF when given the corresponding inputs (Davis, 1999). This allowed Nigrin (1993) to develop a model that could identify words embedded within a continuous input string that does not represent word boundaries (much like continuous speech which does not include straightforward word boundaries). So, rather than recognizing words embedded in a larger pattern by first parsing words at boundaries (for example, attempting to parse spoken words based on stress, phonological bigram frequency, etc.), this model would parse words by first recognizing them. This is actually quite difficult for many networks to achieve because superset words provide the maximum amount of excit-

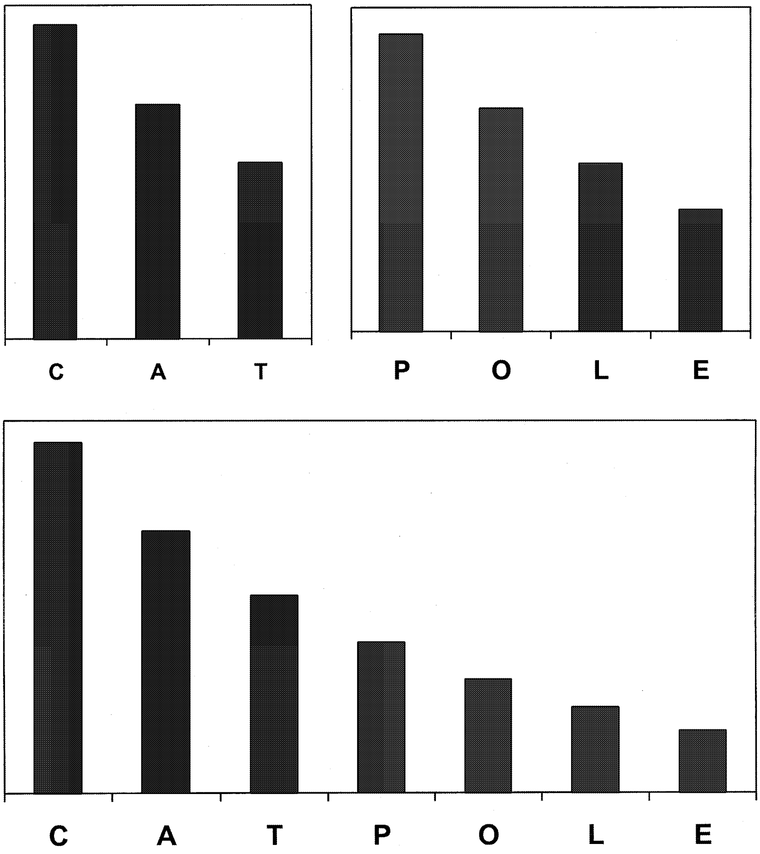

Fig. 1. Spatial coding for morphologically simple and complex words. The order in which letters occur within a word is coded by the relative activity of the letter nodes, with a fixed ratio of activation between adjacent letters. As a consequence, the pattern of activation is the same for POLE presented by itself and within the compound word CATPOLE, although the overall level of activation of the letters in the different contexts differ. The constant pattern of activation allows POLE to be identified in any context.

atory input to both themselves and their subset patterns; for example, SELF activates all the letters for both SELF and ELF (cf. Cohen & Grossberg, 1987).[3] These problems are simply avoided when networks are restricted to words of a given length (e.g., McClelland & Rumelhart, 1981), or mini-

[3] Indeed, Levelt (1989) claimed it was impossible to access whole word forms from component features, regardless of whether the features are localist or distributed, due to what he called the hyperonym and hyponym problem (i.e., problems distinguishing between subsets and supersets). However, the success of models that use spatial coding schemes demonstrate that this conclusion is mistaken. For brief description of a formal proof that sub-set and super-set patterns can be identified, see Bowers (1999).

mized when monosyllable words are presented in isolation (e.g., Harm and Seidenberg, 1998).

Importantly, spatial coding not only supports the identification of complex word forms in complex environments, but it is also consistent with key empirical findings associated with the identification of monosyllable words, some of which pose challenges for more standard coding schemes. For example, spatial coding addresses the so-called *correspondence problem* in visual word identification (Davis, 1999). On the slot-based coding schemes described above, word neighbors that differ from one another by a single letter (e.g., POLE–HOLE) are similar by virtue of sharing three of four input units; namely, *O-in position-2*, *L-in-position3*, and *E-in position-4*. This is appropriate given evidence that word neighbors are perceived as similar within the orthographic system (e.g., Andrews, 1989). However, slot coding does not capture the similarity of another class of words, namely transposed-letter (TL) words that share the same set of letters but with two adjacent letters switched (e.g., CALM–CLAM), as they only overlap in two input units: *C-in-position-1* and *M-in-position-4*. CALM–CLAM are no more similar than CALM–CHUM, for example. This is problematic given evidence that transposed letter words are more similar (and confusable) than orthographic neighbors that differ in only one letter (Andrews, 1996; Chambers, 1979; Forster, Davis, Schoknecht, & Carter, 1987; Taft & van Graan, 1998). Note, on the Wicklecoding scheme CALM–CLAM share no units in common, which predicts that these words are no more similar than CALM–DEAD.

By contrast, spatial coding schemes capture the relative similarity of neighbors and transposed letter words. In the case of the neighbors POLE-HOLE, the same pattern of activation occurs across the letters O, L, and E, and in the case of the transposed letter items CALM–CLAM, the same four letters are activated, and only the relative activation of A and L is switched. Indeed, Davis (1999) has shown that his SOLAR model that employs the spatial coding scheme finds TL-words more difficult to identify than neighbors, as well as replicate many standard findings in the literature.

In sum, connectionist networks with distributed coding schemes cannot yet accommodate the identification of novel morphologically complex words (which is the rule rather than exception in some languages, e.g., Finnish), and it is not clear how the coding schemes employed in these models can be modified to address these problems. At the same time, the spatial coding schemes used in various networks that learn localist lexical representations have already shown some success in accomplishing these and related functions.

5. Language processing beyond single words

Connectionist networks with distributed coding schemes do not support the identification of complex word forms, and the problem becomes more severe when considering how to represent multiple words at the same time. Although a capacity to represent multiple items might seem irrelevant to the task of modeling language input or output (we only perceive or produce one word at a time), the underlying cognitive processes that support language almost certainly encode multiple items simultaneously. For instance, it seems unlikely that our thoughts that drive language are composed as a series of single concepts represented one at a time in sequence. Without the capacity encode multiple co-active items, we can't entertain novel thoughts, such as THE ELK SHOT THE KING. It is a non-starter to code this concept as THE ELK by itself, followed by SHOT (at which point, the thinker would have lost track of THE ELK), followed by THE KING (at which point the thinker does not even know that anyone was SHOT, or that an ELK was involved). Similarly, our phonological systems appears to encode multiple items simultaneously (approximately four according to Cowan, 2001); that is, phonology supports a phonological working memory.

Accordingly, the representations of semantics and phonology will need to support co-active items. To the extent that reading exploits the semantic and phonological systems engaged in language processing more generally (cf. Baddeley, Gathercole, & Papagno, 1998; Martin, Shelton, & Yaffee, 1994), these constraints must also apply to developing models of reading single words as well. In fact, Harm and Seidenberg (1999) designed the phonological component of their model in light of similar considerations:

> Children bring to the reading acquisition task considerable knowledge of phonological structure derived from experience with spoken language. This is an important aspect of the child's experience that previous models have ignored. For example, the architecture of the Seidenberg and McClelland model included a set of phonological units that would allow the network to represent the pronunciations of words, but this representation did not itself encode very much information about the structure of English phonology... In the simulations presented below, we addressed how the existence of prior knowledge of phonological structure—and differences in the quality of this knowledge—affected learning to read. (p. 492).

And indeed the phonological (and orthographic) coding schemes employed by Harm and Seidenberg (1999) are better than the earlier PDP schemes. For instance, the model supports a high degree of accuracy in pronouncing monosyllable nonwords (unlike Seidenberg & McClelland, 1989), provides a more detailed account of reading acquisition and developmental dyslexia compared to Plaut et al. (1996), and provides a preliminary account of phoneme perception, an issue outside the scope of prior PDP models of reading. Nevertheless, if a linkage between the phonological and semantic systems involved in spoken and written language are taken seriously, then it will be important that the phonological representations support a working memory, and the semantic representations support co-active representations (and the relation between these items).

This leads to a serious difficulty as distributed systems can only represent one thing at a time. A pattern of activation across all the units defines a single item, and overlapping two patterns over the same set of units results in a blend that is ambiguous given that there is no way to determine which features belong to which item, the so-called *superposition catastrophe* (von der Malsburg, 1986). So for example, in Fig. 2, it is not possible to determine whether a slightly active node reflects the mixing of large positive activation associated with one item and the smaller negative activation of another, or vice versa. Indeed, this ambiguity applies to all nodes, whatever their activation.

To highlight this limitation, consider a set of models that represent the semantics of individual words using distributed coding schemes (Becker, Moscovitch, Behrmann, & Joordens, 1997; Borowsky & Masson, 1996; Joordens & Becker, 1997; Masson, 1995; McRae et al., 1997; Plaut, 1996; Plaut & Booth, 2000). In these models, concepts are represented by a distributed pattern of activity over a large number of interconnected processing units such that related concepts are represented by similar (overlapping) patterns. One of the achievements of these models is that they can account for various semantic priming effects, such that presenting the prime HAND facilitates the encoding of the related target FOOT more than an unrelated prime CARD (although see Dalrymple-Alford & Marmurek, 1999, for some complications). This occurs because the similar patterns of activation are generated by semantically related words, facilitating the transition from the prime state to the target state when prime-target pair are related. That is, current PDP models exploit the superposition catastrophe in order to account for semantic priming, and as a consequence, priming is obtained at the cost of encoding meaning, such as the simple concepts DOG AND CAT.

Surprisingly, the superposition catastrophe has received little attention in the literature. A brief analysis of this problem was made by Hinton and Shallice (1991) who note in a footnote that mixing the activation patterns

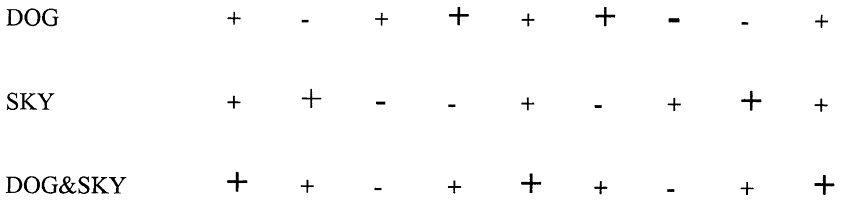

Fig. 2. An illustration of the superposition problem. On top is a distributed pattern of activation for the word DOG presented in isolation, in the middle a pattern for the work SKY presented in isolation, and on bottom a blend pattern of the words DOG and SKY. The blend pattern is ambiguous as it is not possible to determine the activation of the constituent patterns given the blend.

of two items results in a blend pattern that is as close (if not closer) to the activation pattern of each constituent item compared to any other learned pattern, although this was no longer true when three patterns were mixed. A related point was made by Gaskell and Marslen-Wilson (1999) who considered the blends produced from ambiguous word inputs—such as the beginnings of words consistent with a number of word completions, the so-called cohort set. The authors trained a network on a set of items and then presented it with the ambiguous inputs and let the model settle into its final complete state. They demonstrated that completed patterns were often more similar to the cohort items compared to all the non-cohort items, although this property varied considerably depending on the organization of the distributed representations. But as far as I am aware, little additional work has been done concerning the capacity of networks to encode two things at once (for related discussion, see Besner & Joordens, 1995; Kawamoto, Farrar, & Kello, 1994; Masson & Borowsky, 1995).

Although these analyses are interesting, they do not provide a solution to the problem. What these authors have shown is that a blend pattern can—under restricted conditions—be more similar to the representations of the constituent items compared to other items that have *already been presented* to the network. However, this property of distributed systems cannot serve as the basis of a general solution. Perhaps most problematic, blends are not necessarily the product of combining pre-trained patterns. Imagine the situation in which two words are co-active in a distributed phonological system. Although the blend pattern may be more similar to the two constituent words compared to any other trained word, the pattern is not more similar to many *possible* items (or possible blends). The blend pattern might have been produced by combining two nonwords, for example, although this possibility cannot be recovered from the blend. But we can co-encode two novel items: e.g., phonologically, as BLIP–BLAP in short term memory, or conceptually, as "the BLIP is larger than the BLAP"—whatever that means. That is, blend patterns in distributed systems are deeply ambiguous. And in any case, it is not clear what one can do with a blend. Although a blend of two words may be most similar to the coding of the two constituents, the blend pattern is still quite different from each constituent. If a phonological blend is to drive articulation, what would be produced? Without some additional constraints, presumably a blend. Although blends do occur in speech production (e.g., a speaker who co-activated the two words ATHLETE and PLAYER articulated the blend ATHLER; Dell, 1986), they are the exception, not the rule.

In discussing these matters with colleagues a number of solutions have been suggested. One response has been that the superposition catastrophe is only a pseudo-problem, in that the mind may only encode one thing at a time. And indeed, PDP models that code single items have been applied to domains that appear to require co-active representations, including short-term memory. For instance, distributed models of STM have been developed that code for a sequence of items on-line such that each to-be-remembered item is retrieved in turn (for different ways to implement serial retrieval in a distributed system—see Brown, Preece, & Hulme, 2000; Farrell & Lewandowsky, 2002). However, even if one grants that these PDP models do an adequate job in modeling performance on digit span and related tasks, there are other situations in which it is difficult to claim that only a single item is coded at one time. For example, as noted above, it is unclear how a sequence of items could encode the concept THE ELK SHOT THE KING. A behavioral manifestation of co-active representations can be found in a classic form of speech error; namely, lexical exchanges, such as converting the thought "Writing a letter to my mother" into the utterance "Writing a mother to my letter" (Dell, 1986). The standard explanation of this effect is that co-active conceptual representations are assigned the incorrect grammatical roles (in this case, *letter* was assigned the role of indirect object rather than direct object), leading to the exchange. Indeed, according to Dell, Burger, and Svec (1997) "...PDP recurrent network models, which lack separate frame structures [which involve localist coding] ... are not capable of explaining the existence of exchange errors.." (p. 142). Perhaps Dell et al. will be proved wrong, and recurrent networks of sentence processing that encode one word at a time over a set of units (e.g., Elman, 1990; Miikkulainen, 1996; Tabor & Tanenhaus, 1999) will be able to be extended to support this phenomenon, but it is worth noting that no proposal has been advanced thus far.[4] At minimum, if one adopts the view that the superposition catastrophe is a non-issue, then the challenges associated with coding one thing at a time should be acknowledged, particularly when making claims about language (a cognitive skill that appears to require the encoding of co-active representations and putting them in some sort of relation).

A second response has been to agree that the mind can support co-active items in memory, but argue that the superposition catastrophe can be overcome within the PDP framework using "sparsely distributed" coding schemes. Sparsely distributed coding refers to the situation in which each item (e.g., word, object, concept, etc.) is coded by the activation of only a small proportion of units within a bank of many units—for example, including a model with 200 hidden units with only 2 or 3 units active per word.

[4] It is interesting to note that recurrent networks are often more successful in encoding sequences of words when they include localist input coding schemes. When Elman (1988) attempted to model sequential order with distributed representations, he wrote that the: "network's performance at the end of training... was not very good" After five passes through 10,000 sentences, "the network was still making many mistakes" Elman (p. 17). Much greater success was obtained when he relied on localist coding schemes (Elman, 1990). It continues to be the case that most recurrent networks of sentence processing tend to use localist word coding schemes (e.g., Tabor and Tannhaus, 1999; but see Miikkulainen, 1996).

Note, critical to the concept of sparse *distributed* coding is that an item is still defined as a pattern of activation over multiple units, and that each unit contributes to the coding of multiple items (such that no-one unit can uniquely define a piece of knowledge)—otherwise, this concept is not logically distinct from localist coding schemes.

Although this proposal seems intuitively plausible, the only research relevant to this question suggests just the opposite—namely, that the superposition catastrophe becomes more serious with sparse coding (Gaskell & Marslen-Wilson, 1999). As noted earlier, these authors considered the word blends produced from ambiguous word inputs—such as the beginnings of words consistent with a number of word completions (the cohort set). They demonstrated that completed patterns within a distributed system were often more similar to the cohort items compared to all other items, although they noted that this property varied considerably depending on the organization of the distributed representations. The condition in which the blend patterns best distinguished between cohort and non-cohort items was when the coding scheme was the least sparse. They write "This argues against any attempt to improve the ability of a distributed system to coactivate words by making representations sparser, or near-localist. Any reduction in the overlap between word representations comes at the cost of increased interference between coactive representations" (p. 449). This is a striking contrast to the lack of interference reported between co-active localist codes.

In any case, sparse coding appears a poor format for coding word knowledge. One advantage of sparse coding within the PDP framework is that it is relatively immune to "catastrophic interference" in which new learning erases old (McClelland, McNaughton, & O'Reilly, 1995). But this advantage has an associated cost: namely, it supports poor generalization. For these and related reasons sparse coding has been suggested as a possible medium for episodic memory (where resistance to interference is critical), with more fully distributed coding schemes employed within associated systems in order to support generalization (McClelland et al., 1995). Key for the present context is that models of word identification require a coding scheme that supports generalization in order to name novel words (i.e., nonwords). So, even if additional analysis reveals that sparse coding can indeed overcome the superposition catastrophe under certain conditions, it is not clear the solution would be appropriate in the context of building a model of reading. And if a distributed coding scheme is discovered that supports generalization and overcomes the superposition catastrophe, it will also be important to show that it can encode relational information amongst co-active representations (e.g., order amongst a set of words)—as is currently achieved with localist models that employ spatial coding, as discussed below.

There is of course one way in which PDP models with distributed representations can support co-active representations: each item can be coded in non-overlapping banks of units. So for example, the novel concept THE

ELK SHOT THE KING might be coded as a distributed pattern of activation over three banks of units, with one distributed pattern encoding THE ELK, a second distributed pattern of activation over a separate bank of units coding SHOT and another distributed pattern over a third set of units coding for THE KING. More generally, different banks of units might be reserved for the different constituents of a thought, with THE ELK coded across a set of units reserved for the *agent* role, SHOT encoded across a set of units reserved for an *event* role, and THE KING coded across a set of units reserved for the *recipient* role of a thought. In this way, the novel concept THE SQUIRE KISSED THE ELK could be coded as different patterns of activation across the same agent, action, and recipient units. This is the general approach adopted by Hinton (1986), Rumelhart and Todd (1993), St. John and McClelland (1990; St. John, 1992), and others in order to encode relations between multiply co-active items (even when employing localist coding, e.g., St. John & McClelland, 1990).

But there are problems with this solution. A first issue, although perhaps not critical, is that this approach may undermine some of the past successes of PDP models taken to support distributed representations. For instance, as noted earlier, a number of authors have taken advantage of the blend patterns within a common bank of units in order to explain semantic priming. That is, the transition from one semantic pattern to another was facilitated for related compared to unrelated words. However, if different words are coded on non-overlapping banks of units, then the similar activation patterns of related words would not impact on performance—DOCTOR would be coded in one place, NURSE in another. Similarly, a number of PDP models of semantics have employed a common bank of semantic features in order to explain various semantic disorders (e.g., Farah & McClelland, 1991)—with the success of the models taken as evidence in support of distributed coding schemes. But if indeed there are separate banks of units involved in coding for agents, actions, recipients, etc., then the relevance of these findings are unclear. Indeed, a semantic network with multiple banks of non-overlapping units would appear to make some surprising predictions. For instance, it should be possible to find a person with a semantic disorder who has difficulty in conceiving of a DOG in some contexts (e.g., The DOG liked the CAT) but not others (e.g., the CAT liked the DOG). Although there are forms of anomia in which a person has difficulty in naming nouns (a watch) but not verbs (to watch), and vice versa (e.g., Caramazza & Hillis, 1991), I am not aware of examples of patients who can conceive JOHN LOVES MARY but not MARY LOVES JOHN.

More importantly, the required number of non-overlapping banks of units would need to scale directly with the complexity of the thought. For instance, in order to encode JOHN AND MARY LOVE DOGS AND CATS two sets of units are needed in the agent position and another two in the recipient position, with the function of distributed coding unclear.

By contrast, if localist coding is employed there is no need to include multiple banks of units—DOG and CAT could simply be co-active within the same bank of units. Given the high cost of employing distributed coding schemes in this context, their value should be made explicit.

And critically, by including non-overlapping banks of units for the different constituents in a thought (e.g., agent, action, and recipient), the same concepts are treated as unrelated in the different contexts. So if a person is informed that JOHN TALKED TO JACK, he/she is in no position to know whether JACK TALKED TO JOHN, as JOHN the agent (JOHN1) is unrelated to JOHN the recipient (JOHN2), and JACK2 is unrelated to JACK1 (Marcus, 1998, would say that JOHN2 and JACK1 are outside the models training space). To be more concrete, consider the network schematized in Fig. 3 that depicts a pattern of activation over three banks of units. The network could learn to reinstate all three patterns given two patterns (e.g., given JOHN1 and TALK the network could infer JACK2), learn to code many other facts (e.g., PETER1 KISSED JANE2), as well as make some forms of generalization (e.g., Hinton, 1986; St. John & McClelland, 1990). For example, after learning various facts about JOHN1, a network, under the appropriate conditions, can make some inferences about JOHN1 (e.g., the network might activate the appropriate units in the recipient bank of units when JOHN1 and KISS are activated—despite the fact that JOHN1 had not kissed anyone before). However, there is no way for the model to infer JACK TALKED TO JOHN having learned that JOHN TALKED

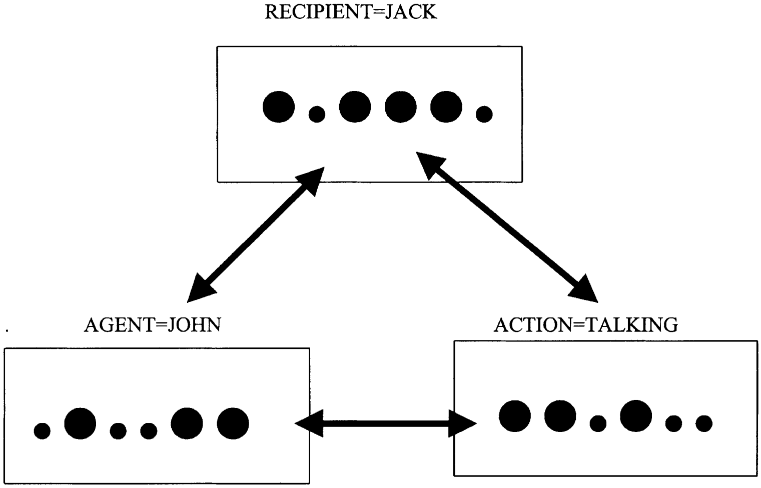

Fig. 3. An illustration of how JOHN as agent, JACK as recipient, and TALK as action could be coded as a distributed pattern of activation over three banks of units. While this pattern codes for the concept JOHN IS TALKING TO JACK and a different pattern across the same units might code for THE DOG IS CHASING THE CAT.

TO JACK (the model has no experience with JACK1 nor JOHN2). The model could activate JACK2 given JOHN1 and TALK, but this the old sentence JOHN TALKED TO JACK. Again, the problem is that all the items are coded in a context dependent fashion—e.g., John-as-agent, John-as-recipient. Context independent representations (e.g., JOHN1 = JOHN2, JACK1 = JACK2) appear necessary for various sorts of inference, including systematic thought, as in the above example (Fodor & Pylyshyn, 1988).

In sum, a key limitation of distributed coding schemes is that they can only encode one item at a time over a common set of units (unambiguously). Accordingly, most PDP models of word naming (e.g., Harm & Seidenberg, 1999), phonological working memory (Brown et al., 2000), and semantics (e.g., Plaut & Booth, 2000) are restricted to processing one item at time. The only exception are models that include separate banks of units for each co-active word, a solution with its own problems, as noted above. Of course, it may turn out that the superposition catastrophe is only a pseudo-problem (i.e., the mind only encodes one word at a time) or it may be a real problem that can be overcome while maintaining distributed lexical coding (e.g., sparse coding). But given the strong commitment to distributed lexical knowledge within the PDP camp, advocates of distributed representations need to confront this issue raised by von der Malsburg (1986) over 15 years ago.

5.1. Possible solution involving localist coding schemes

Although the superposition catastrophe poses a serious challenge for PDP models, it is important to emphasize that there are connectionist models of semantics, phonology and word identification that both learn and represent multiple pieces of information simultaneously and on-line (in a context independent fashion), but they all rely on localist representations.

For example, in the case of semantic knowledge, Hummel and Holyoak (1997) developed a model capable of learning and representing complex thoughts by relying on localist concept units. The prime function of the localist representations was to bind together the semantic features associated with a concept, thus avoiding ambiguous blends. So for example, the concept DOG might include the (localist) features PET, BARK, TAIL, 4-LEGS, and the concept CAT the features PET, PURR, WHISKERS, 4-LEGS, with the features bound to localist representations DOG and CAT, respectively. Critically, the features (and concepts) are coded in a context independent fashion such that the same units code for the same information in all situations. For instance, the concept DOG is coded by the same unit regardless of whether DOG is the agent or recipient of a thought. Similarly, the concept PET and TAIL are coded by the same units, regardless of whether they are bound to DOG or CAT (allowing the model to represent the similarity of DOG and CAT). By performing mappings over these context independent units, the model can support systematic thought (e.g., inferring that JACK TALKED TO JOHN if the model was taught that JOHN TALKED TO JACK), among other key capacities beyond current PDP models.

In the case of phonology, Page and Norris (1998) developed a localist connectionist model of phonological working memory in which word order is coded in terms of the relative activation of co-active localist word representations—so for instance, if their model was presented with the list of words DOG, CAR, SUN, and FUN to remember, then all four words become co-active, with DOG the most active, and a constant decreasing gradient of activation across items, with items retrieved from short-term memory in the order of activation—what the authors called a primary gradient. The authors demonstrate primary gradient models can accommodate a wide range of the STM phenomenon in the literature. And although this particular model did not learn its representations, there are primacy gradient models of working memory that do (e.g., Nigrin, 1993).

Although these issues may seem well removed from developing models of word naming, many of the same constraints (and solutions) apply to the task of naming and identifying words (and objects; see Hummel & Biederman, 1992). For example, the primary gradient employed in the working memory model of Page and Norris (1998) is the very same as the spatial coding schemes introduced by Grossberg (1978), and applied to localist letter units in order to identify spoken and written words (Nigrin, 1993; Davis, 1999). To take the Davis (1999) model, a localist letter coding scheme allowed it to support the co-activation of multiple letters, and a localist word coding scheme supported the co-activation of multiple word units. And the spatial coding applied to the localist letter and word representations allowed the model to encode the relative order of letters in a context independent fashion—that is, the same D unit is involved in coding DOG and WORD. This in turn allowed the model to identify POLE in the novel context CATPOLE (as in Fig. 2), just as the context independent semantic representations included within the Hummel and Holyoak (1997) model allowed it to identify the equivalence of JOHN in JOHN LOVES MARY and MARY LOVES JOHN.[5]

Note, in all cases, improved generalization was achieved by coding knowledge (e.g., letters, words, concepts) in a context independent fashion, and by operating over these abstract categories. That is, these model all im-

[5] Spatial coding is not sufficient in the case of semantics because the possible relations between co-active representations are more diverse than that of order (i.e., which letter comes first, second, etc.). As a consequence, Hummel and Holyoak (1997) employed another mechanism of coding relations in a context independent fashion, namely synchronous firing of units. Also see Shastri and Ajjanagadde (1993). Interestingly, the binding mechanism proposed by Hummel and Holyoak (1997) is limited to encoding approximately four items at a time, consistent with the data described by Cowan (2001). See Hummel and Holyoak (1997) for details.

plement rules (for detailed discussion, see Marcus, 2001). Clearly then, these models provide a radical alternative to the standard PDP models that reject localist coding at a lexical level, and context independent representations at all levels (and rules more generally). Of course, the success of models that learn localist-context-independent representations will need to be further tested and explicitly contrasted with the more familiar PDP approaches before any firm conclusion are warranted. But this process is only hindered when distributed representations (and the absence of rules) are listed as one defining feature of connectionism.

6. Overall summary

The key theoretical contribution of connectionist models is not that they provide a means to accommodate various cognitive phenomena while rejecting all high-level localist representations (e.g., words), but rather, that they have introduced a fundamental set of learning and processing principles that apply broadly to many cognitive domains (cf. Grossberg, 1980; McClelland, Rumelhart, & the PDP Research Group, 1986). And by building models from basic principles, theories are more strongly constrained than abstract computational models that are often hand-tailored to explain a narrow range of findings. Indeed, by developing theories based on a small number of general principles, these models may help to explain why the brain/mind has adopted particular solutions as opposed to other possible solutions that are descriptively adequate—that is, connectionist models can provide explanatory theories (cf. Seidenberg, 1993a).

Although localist or distributed representations can be learned within this framework, there are two general reasons why a strong commitment to distributed coding schemes is unwarranted. First, and most importantly, current models that include distributed coding fail to support a wide variety of cognitive functions. The greatest success of distributed coding schemes has been in the domain of naming and identifying monosyllable words and nonwords, but even within this restricted domain serious empirical and computational challenges remain (e.g., the Plaut et al., 1996 model accounts for less than 1% of nonword naming variance; Spieler & Balota, 1997). The most serious problem, however, is that the models have not been applied to more complex cognitive phenomena that are fundamental to reading and cognition more generally, such as identifying complex word forms, or coding for two words at the same time. Until some proposals are offered as how these more complex skills can be accomplished within this framework, there is little reason to strongly endorse the claim that all lexical knowledge is coded in a distributed format.

The second reason why a commitment to distributed coding schemes is unwarranted is that, unlike the common view, connectionist models can learn localist representations (i.e., the representations do not need to be stipulated), and they show promise in supporting many of the functions on which distributed systems fail. In particular, these models can support the identification of complex word forms (e.g., Davis, 1999; Nigrin, 1993), can generalize in systematic fashions (e.g., Hummel & Holyoak, 1997), map easily between arbitrary domains (e.g., Carpenter & Grossberg, 1987), can represent multiple items simultaneously (e.g., Cohen & Grossberg, 1987; Nigrin, 1993), and can represent order amongst a set of items in short-term memory (e.g., Bradski et al., 1994; Page & Norris, 1998). And as shown by Page (2000), the standard criticisms levied against localist coding schemes are unwarranted.

It is also important to emphasize that models with localist lexical representations support a wide range of skills using a small set of principles, one of the important goals of the connectionist agenda. Indeed, Grossberg has developed ART and related networks around a small set of key functional demands, including: (a) the ability to process information in noise, such that a network can encode significant events when the input to the system is small or large; the so called *noise-saturation dilemma*, (b) the ability to learn new information without erasing past knowledge without artificially constraining the nature of the learning environment, the so-called *stability–plasticity dilemma* (Grossberg, 1976, more commonly termed catastrophic interference, McCloskey & Cohen, 1989; Ratcliff, 1990);[6] (c) the ability to learn in real time with or without a teacher, (d) the ability to identify, learn and recognize subset and superset problems, the so-called *temporal chunking problem*, (e) the ability to learn with only local interactions, such that learning is physiologically plausible (unlike back-propagation), and (f) as noted above, the ability to represent order amongst multiply active items. All these functions have been met employing a small number of general principles that apply to a wide range of cognitive phenomena, as noted above. Before localist coding schemes are rejected, models with distributed coding schemes should match this performance.

7. Concluding comment

One possible reaction to the argument I've put forward is that it is not surprising that connectionist models that reject localist lexical knowledge are limited, and that future developments may well address these outstand-

[6] Note, the ART solution to the stability–plasticity dilemma undermines a key claim made by McClelland et al. (1995). The authors listed catastrophic interference as one of the three principles of connectionist learning. In particular, Principle 2 was described as: "Attempts to learn new information rapidly in a network that has previously learned a subset of some domain lead to catastrophic interference" (p. 435). But ART can learn new information rapidly without interference.

ing problems. Supporters of the distributed approach might argue that you have to start somewhere, which is certainly true.

But the suggestion that modelers must first address simple phenomena before tackling more complex problems is, in some ways, a mischaracterization of the approach adopted by many advocates of the PDP framework. The current limitations of connectionist networks with distributed representations are fundamental—the most complex phenomena considered in any detail was the capacity to encode a relation of two co-active items (e.g., DOG and FISH in semantics, or co-activating CAT and POLE within orthography when given the novel compound CATPOLE). At the same time, PDP models have already been applied to complex patterns of experimental data concerning naming, lexical decision, and semantic priming for single syllable items. One of the consequences of this approach is that authors have developed models of semantic memory that support a complex pattern of semantic priming at various SOAs but which cannot represent meaning.

Quite different models are developed when one takes an alternative tack. Rather than initially focusing on the experimental details within a restricted domain, one might ask more basic computational (engineering) questions, of the sort mentioned above: How to learn new information without erasing old information (within limits), how to represent order amongst co-active items, how to represent sub- and super-set patterns, how to learn with and without an external teacher, etc. By first considering these most fundamental questions one may be in the position to identify basic constraints on network architectures, and then build on these foundations by relying on the detailed experimental results reported in the literature. Indeed, this has been the approach of Grossberg and colleagues (amongst others), which has lead to very different network architectures, including networks that often rely on localist coding schemes.

Despite the computational power and neural plausibility of connectionist networks that learn localist representations, such as the ART models of Grossberg and colleagues, this work is almost entirely ignored in the psychological literature. Although there are a few examples of researchers associated with the PDP approach quoting Grossberg (e.g., Stone & VanOrden, 1994), the number of references to ART and related models by PDP advocates is not far from zero. The field needs to more fully consider existing localist approaches, and explicitly contrast models that learn localist and distributed codes. Only then can strong conclusions regarding the nature of learned knowledge be advanced.

Acknowledgments

Jeffrey S. Bowers, Department of Experimental Psychology. I would like to thank Patricia Bowers, Markus Damian, Colin Davis, John Hummel, Stephan Lewandowsky, Sven Mattys, Kate Nation, Keith Rayner, and four anonymous reviewers for their helpful comments.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Memory & Cognition, 15*, 802–814.

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language, 35*, 775–800.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158–173.

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning Memory and Cognition, 23*, 1059–1082.

Besner, D. (1999). Basic processes in reading: multiple routines in localist and connectionist models. In R. M. Klein & P. A. McMullen (Eds.), *Converging methods for understanding reading and dyslexia*. Cambridge, MA: MIT Press.

Besner, D., & Joordens, S. (1995). Wrestling with ambiguity—further reflections—reply. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 515–519.

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

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*, 63–85.

Bowers, J. S. (1999). Grossberg and colleagues solved the hyperonym problem over a decade ago. *Behavioral and Brain Sciences, 22*, 38.

Bradski, G., Carpenter, G. A., & Grossberg, S. (1994). Store working-memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics, 71*, 469–480.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107*, 127–181.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349*(6312), 788–790.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37*, 54–115.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks, 4*, 565–588.

Chambers, S. M. (1979). Letter and order information in lexical access. *Journal of Verbal Learning and Verbal Behavior, 18*, 225–241.

Cohen, M., & Grossberg, S. (1987). Masking fields: A massively parallel neural architecture for learning, recognitzing, and predicting multiple groupings of data. *Applied Optics, 26*, 1866–1891.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of sematnic processing. *Psychological Review, 82*, 407–428.

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

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.

Dalrymple-Alford, E. E., & Marmurek, H. H. C. (1999). Semantic priming in fully recurrent network models of lexical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 758–775.

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

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*, 123–147.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production—a comment. *Psychological Review, 98*, 604–614.

Elman, J. L. (1988). Finding structure in time (CRL Technical Report 8801). La Jolla, CA: UCSD.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–216.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment—modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120*(4), 339–357.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 19*, 59–79.

Forster, K. I. (1994). Computational modeling and elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1292–1310.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *Quarterly Journal of Experimental Psychology, 39A*, 211–251.

Fodor, J. A. (2000). *The mind does't work that way*. Cambridge, MA: MIT Press.

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

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition and blending in spoken word recognition. *Cognitive Science, 23*, 439–462.

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

Grainger, J., & Jacobs, A. M. (1998). On localist connectionism and psychological science. In J. Grainger & A. M. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition* (pp. 1–38). Mahwan, NJ: Lawrence Erlbaum Associates.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics, 23*, 187–203.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review, 87*, 1–51.

Grossberg, S. (1978). A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Psychological Review* (87,. *Progress in theoretical biology (pp. 233—374)* pp. 1–51). New York: Academic Press.

Grossberg, S. (1987). Competitive learning—from interactive activation to adaptive resonance. *Cognitive Science, 11*, 23–63.

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Consciousness and Cognition, 8*, 1–44.

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

Hinton, G. E., McClelland, J. L., Rumelhart, D. E. (1986) Distributed representations. In: D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. explorations in the microstructure of cognition.* Vol. 1. *Foundations* Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Leasoning an attractor network—investigations of acquired dyslexia. *Psychological Review, 98*, 74–95.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape-recognition. *Psychological Review, 99*, 480–517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427–466.

Joordens, S., & Becker, S. (1997). The long and short of semantic priming effects in lexical decision. *Journal of Experimental Psychology: Learning Memory and Cognition, 23*, 1083–1105.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When 2 meanings are better than one—modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1233–1247.

Levelt, W. J. M. (1989). *Speaking: From intetion to articulation*. Cambridge, MA: MIT Press.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*, 243–282.

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA.

Masson, M. E. J. (1995). A distributed-memory model of semantic priming. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 3–23.

Masson, M. E. J., & Borowsky, R. (1995). Unsettling questions about semantic ambiguity in connectionist models—comment. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*(2), 509–514.

Martin, R. C., Shelton, J. R., & Yaffee, A. (1994). Language processing and working memory: Neuropsycholgical evidence for separate phonologial and semantic capacities. *Journal of Memory and Language, 33*, 83–111.

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

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3*, 166–168.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. 1.an account of basic findings. *Psychological Review, 88*, 375–407.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (1986). *Parallel distributed processing: Psychological and biological models* (Vol. 2). Cambridge, MA: MIT Press.

McClelland, J. L., & Seidenberg, M. S. (2000). Why do kids say goed and brang? Review of S. Pinker, Rules and Words. *Science, 287*, 47–48.

McLeod, Plunkett, & Rolls, (1998). *Introduction to connectionist modeling of cognitive processes*. New York, NY: Oxford University Press.

McRae, K., deSa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*, 99–130.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation*. New York: Academic Press.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science, 20*, 47–73.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing models of visible language* (pp. 259–268). New York: Plenum.

Murre, J. M. J., Phaf, H., & Wolters, G. (1992). CALM: Categorizing and learning module. *Neural Networks, 5*, 55–82.

Nigrin, A. (1993). *Neural networks for pattern recognition*. Cambridge, MA: MIT Press.

Page, M. P. A. (1994). Modeling the perception of musical sequences with self-organizing neural networks. *Connection Science: Journal of Neural Computing, Artificial Intelligence and Cognitive Research, 6*, 223–246.

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*, 443–512.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761–781.

Pinker, S. (1999). *Words and rules: The ingredients of langauge*. London: Weidenfeld & Nicolson.

Pinker, S., & Prince, A. (1988). On language and connectionism—analysis of a parallel distributed-processing model of language-acquisition. *Cognition, 28*(1–2), 73–193.

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

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

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

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

Plaut, D. C., & McClelland, J. M. (2000). Stipulating versus discovering representations. *Behavioral and Brain Sciences, 23*, 489–491.

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

Plaut, D. C., & Shallice, T. (1993). Deep Dyslexia—a case-study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*, 377–500.

Ratcliff, R. (1990). Connectionist models of recognition memory—constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285–308.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Foundations* (Vol. 2). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. *Attention and Performance, 14*, 3–30.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9*, 75–112.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic binding using temporal asynchrony. *Behavioral and Brain Sciences, 16*, 417–494.

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

Seidenberg, M. S. (1993b). A connectionist modeling approach to word recognition and Dyslexia. *Psychological Science, 4*, 299–304.

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. *Psychological Review, 97*(3), 447–452.

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9*, 234–237.

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

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411–416.

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

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–457.

Stone, G. O., & VanOrden, G. C. (1994). Building a resonance framework for word recognition using design and system principles. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1248–1268.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science, 23*, 491–515.

Taft, M., & van Graan, F. (1998). Lack of phonological mediation in a semantic categorization task. *Journal of Memory and Language, 38*, 203–224.

von der Malsburg, C. (1986). Am I thinking assemblies?. In G. Palm & A. Aertsen (Eds.), *Brain theory*. Berlin: Springer.

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 50a*, 439–456.

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