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

Jeffrey S. Bowers
Department of Experimental Psychology
University of Bristol
Bristol, England
BS8-1TN
(+44) (0)117 928-8573
j.bowers@bris.ac.uk

Running head: Localist vs. distributed representations

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 within the connectionist camp should be considered a specific theoretical claim regarding the structure of knowledge rather than a core principle, as often assumed. In this article, 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. It is concluded that the common rejection of localist coding schemes within connectionist architectures is unwarranted.

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

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 some other plausible alternatives; that is, this approach shows promise of supporting explanatory rather than descriptive theories (Seidenberg, 1993a).

One claim often associated with this framework is that knowledge is coded in a distributed fashion. That is, information is coded as a pattern of activation across many processing units, with each unit contributing to many different representations. Indeed, at this point, the concepts connectionism and distributed representations are so strongly associated that distributed representations are sometimes described as one of the core "connectionist principles" (e.g., Seidenberg, 1993b, p. 300). This challenges one of the fundamental assumptions of more traditional "symbolic" approaches to cognitive theorizing according to which knowledge is coded in a localist format, with separate representations coding for distinct pieces of information.

The problem with this claim, however, is that localist representations can be learned in connectionist architecture (e.g., Carpenter & Grossberg, 1987; Grossberg, 1980) and many connectionist models depend on localist coding for their functioning (e.g., Dell, 1986; Grainger & Jacobs, 1996; Grossberg, 1985; Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Page & Norris, 1998; see Grainger & Jacobs, 1998, for a recently edited book on "localist connectionism"). So it is clearly a mistake to describe distributed representations as an intrinsic property of connectionism. 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 knowledge is represented in a distributed format (if ever), and under what conditions localist representations develop (if ever).

The main goal of this paper is to raise these questions and put forward a set of challenges that need to be met before a commitment to distributed coding schemes is warranted. Page (in press) 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, I challenge past evidence that has been taken in support of distributed coding schemes, question the functional utility of these codes, and most importantly,

identify fundamental computational limitations of all current models that reject localist representations. Many of these challenges are raised in the context of theories of language processing (in most cases reading), although these issues are discussed in other domains as well. At the same time, and in contrast with these limitations, I show that connectionist models that learn localist representations have met many of these challenges, although these successes have largely been ignored in the psychological literature.

Background Issues

Before raising any specific challenges, 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 model 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 models can be excluded from this category. This seems to be the position of Seidenberg (1993b) when he calls distributed representations a core connectionist principle, allowing him to exclude these and more recent models, such as those by Dell (1986), Grainger and Jacobs (1996), and others, which are also hand wired with localist representations.

But even if one agrees that learning is one of the defining features of connectionist models, the critical point that needs to be emphasized is that connectionist models can learn localist representations. This, however, is not widely recognized. For example, consider the following comment by an anonymous review of an earlier draft of this paper:

> The most important problem is that the kind of localist models that both authors favor [That is Page and myself] involve stipulating on some pretheoretical or intuitive basis what the units are to represent . That is, someone decides that units should correspond to words, syllables, phonemes, or whatever, and then erect a model….. The present author doesn't acknowledge this issue, which is a central one….. Localist models of this sort <u>all</u> [underline added] solve the representations by fiat: people just invent them. Given the lack of constraints on what kinds of representations can be hand-wired…. this doesn't end up being a very strong criterion."

Similar points have been published elsewhere.

But 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 <u>a priori</u> (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" levels,

and argued that it is preferable to employ the more abstract terms "item" and "list" levels to refer to units extracted by the learning procedure given that the type of unit is not pre-specified in his models. A list level codes 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 have been extended (as well as implemented) by Nigrin (1993) and, with specific focus on visual word identification, Davis (1999). It is simply not true that localist representations need to be hand-wired on some pretheoretical basis.

A second general issue of possible confusion is with regards to the types of localist representations that are rejected by the PDP camp. The defining feature of a localist representation is simple enough: localist coding schemes include separate representations (that is 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 otherwise—is claimed to be distributed within the connectionist framework. Indeed, within the PDP camp, the extent to which models include distributed representations varies considerably. 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 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 et al., 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 (in press), the willingness to include localist sub-lexical codes within a network while strongly rejecting localist lexical representations is surprising, as it leads to the situation in which complex graphemes such as TCH are represented with localist representations whereas high-frequency words such as IT are not. But in any case, the central claim of the present article is that the rejection of localist representations at the lexical level leads to serious limitations that have not yet been confronted by the PDP camp.

The above considerations should also make it clear that there is no more or less stipulation involved in developing models with localist vs. distributed representations. Models that use back-propagation (among some other learning

algorithms) learn distributed lexical representations, whereas most models that learn via adaptive resonance principles (e.g., Carpenter & Grossberg, 1987), and various competitive learning schemes (e.g., Rumelhart & Zipser, 1985) learn localist lexical codes. There is no reason to conclude that one learning algorithm involves stipulation whereas the other does not. Furthermore, as noted above, even in models that learn with back-propagation, authors sometimes stipulate localist grapheme and phoneme units (Plaut et al., 1996), or localist letter and phonetic feature units (Harm & Seidenberg, 1999), or stipulate distributed input and output codes (Seidenberg & McClelland, 1989). Distributed representations are every bit as much "stipulated" as localist representations.

One final point of general introduction should be noted. One of the compelling features of the connectionist models that learn distributed representations is that the same set of principles can be used to explain to a wide range of phenomena, from reading regular and irregular single syllable words (Seidenberg & McClelland, 1989), nonwords (Plaut et al., 1996), as well as various developmental and acquired disorders of reading (e.g., Plaut et al., 1996; Harm & Seidenberg, 1999), semantic priming phenomena (Plaut & Booth, in press), not to mention various non-language phenomena. 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 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). 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 linguist 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. Indeed, no one has attempted to explain a wider range of cognitive phenomena than Grossberg and colleagues, who 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 embedded within ART networks that learn localist representations (cf., Grossberg, 1999). Others have developed and extended this work by constructing models that learn to segment familiar patterns embedded in larger patterns, as is necessary in speech segmentation (Nigrin, 1993) and visual word identification (Davis, 1999) based on some general learning and processing principles first identified by Grossberg. The point is not that Grossberg's general approach is correct (although I confess I find this work very compelling), but rather, that connectionist models with

distributed representations should not be preferred because of their success in a variety of domains. Indeed, as argued below, models with localist codes accomplish a good deal more than current models with distributed representations.

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 lexical representations, modelers must "stipulate" various features of their networks, but this applies equally to models that learn localist and distributed representations, and both approaches have been applied to a wide range of phenomena. Furthermore, as shown by Page (in press) 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 more traditional localist representations, it is not sufficient to appeal to any of these general considerations. Rather, it is necessary to demonstrate the advantages of distributed coding schemes when the two approaches are directly compared in terms of their capacity to accommodate data as well as their ability to scale up to more realistic settings. This is approach taken in the present article, and below I identify a series of basic cognitive functions that have yet to be supported by connectionist models that learn distributed representations, and contrast these limitations with existing connectionist models that learn localist representations.

Identifying and naming single syllable words and nonwords.

Connectionist models of word identification that include distributed representations have focused on the processing of single syllable words and nonwords, with particular emphasis on the mappings between orthography and phonology – conditions particularly well suited for learning distributed representations. In many respects, these models have been successful within this domain. For example, these models can name nonwords (e.g., blap) and irregular words (e.g., pint). 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. In addition, as noted by Plaut (1999) among others, these 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). Let me briefly note two limitations that Besner does not mention. First, 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.

Second, Bowers and Michita (1998) reported evidence that words written in the orthographically unrelated Hiragana and Kanji scripts of Japanese map onto

common orthographic representations, much like upper- and lower-case words in English (e.g., READ/read; e.g., Bowers, Vigliocco, & Haan, 1998; Coltheart, 1981; McClelland, 1976). In particular, we observed robust long-term priming between study and test words written in the Kanji and Hiragana scripts (and vice versa), whereas little or no priming was obtained following a study-test modality shift, suggesting that the cross-script priming was mediated by modality-specific orthographic representations. This finding is consistent with various reports of robust cross-case and modality specific priming in English (e.g., Bowers, 1996). But unlike upper- and lower-case words, Hiragana and Kanji words do not share any sub-lexical representations and can only be mapped together at the whole word level. Similar robust priming effects have been obtained between Hindi and Urdu in Indian (Brown, Sharma, & Kirsner, 1984) and Roman and Cyrillic in Serbo-Croatian (Feldman, & Moskovljevic, 1987), although these latter studies did not demonstrate that the cross-script priming was also modality specific, leaving open the possibility that phonological or semantic representations supported these latter effects (for general review of the priming literature, see Bowers, 2000). The Japanese results, however, support the conclusion that words are represented at a lexical level within the orthographic system, and thus provide a challenge to connectionist theories that reject this level of representation.

Let me also mention one of the most striking problems noted by Besner (1999) in his review; namely, these models account for little variance in reading response times at the item level, despite their success in pronouncing words (and more recently nonwords). 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 were compared to the human data, the model accounted for 10.1% of the variance. By contrast, the combined influences of log frequency, neighborhood density (Coltheart's N), and word length was 21.7% of the variance, showing that there is 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. 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. In addition, Rastle and Coltheart (1997) tested the Plaut et al. (1996) model on the nonwords included in Weeks' (1997) experiment, and found that it only accounted for 1% of nonword naming latency variance, whereas the Coltheart et al. (1993) dual-route model explained 38% of the RT variance (see Seidenberg & Plaut, 1998, for response). Thus, although these models can name single syllable words and nonwords, there is little evidence that they accomplish this in the way humans do.

As emphasized by Besner, future advances in connectionist models with distributed representations may overcome some and perhaps all of these limitations. The important point to note for present purposes, however, is that even if later versions of these models can be modified to accommodate all of these results, the conclusion that distributed representations underlie reading (and cognition more generally) would still be unsupported. In order to justify these strong claims, it is necessary to demonstrate that these representations can support more complex language functions– such as naming two syllable words, morphologically complex

words, or encoding two separate words at the same time. Note, the Plaut et al. (1996) and Harm and Seidenberg (1999) models cannot address these questions as their input and output coding schemes were designed to only represent single syllable items presented one at a time. Still, it is interesting (and perhaps telling) that the key change introduced to the Plaut et al. (1996) model in order to support the pronunciation of single syllable nonwords was the inclusion of localist sub-lexical representations.

In sum, a number of empirical results obtained with single syllable words and nonwords remain to be explained with network models that learn distributed representations, including phenomena that are better accommodated by models with localist representations (cf. Besner, 1999). And in any case, general claims concerning the structure of mental representations based on the processing of simple word and nonwords are unwarranted.

Identifying and naming two syllable words and morphologically complex words.

As noted above, one of the important limitations of current connectionist models with distributed representations is that they cannot support the identification of two syllable words, let alone compound words, such as toothache. 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 these more complex word forms.

However, not only are the input and output coding schemes employed by current models restricted to the processing of single syllable items, there are reasons to think that these approaches will have difficulties scaling up to more 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 & McClellend, 1989) or a slot based approaches in which letters or graphemes are coded in long-term memory in terms of their location within a word (e.g., Harm & Seidenberg, 1999; Plaut et al., 1997). For example, the wickelcoding scheme used by Seidenberg and McClellan (1989) used relational coding, in which each letter is coded relative to their immediate context, so that the word DEAD would be coded as the wickelfeatures #DE, DEA, EAD, AD#, where # refers to a word boundary. The word DEAD would be identified when all the relevant features were identified. In the slot based approach by Plaut et al. (1996), however, DEAD would be represented by the graphemes D, EA, and D in positions one, two, three, respectively (a similar slot-based approach was used by Harm & Seidenberg, 1999).

One problem with both these coding scheme, however, is that they obscure letter-phoneme correspondences that need to be learned through experience. So for instance, the mapping between D -> /d/ is relatively constant across letter positions within a word, but this regularity is lost in the wickelcoding scheme as the two Ds in DEAD are represented by the unrelated orthographic forms #DE and AD#, and this is also true of slot-based approaches, as the two Ds are represented by the units D-in-first-position and D-in-third-position. That is, the regularities that need to be learned are dispersed across unrelated orthographic forms. Plaut et al. (1996) attributed the poor nonword naming performance of the Seidenberg and McClelland (1989) to the

extreme version of the dispersion problem in the wickelcoding scheme, which was reduced in the Plaut et al. (1996) model, and further reduced in the Harm and Seidenberg (1999) model. The reduced dispersion problem in both cases allowed these latter models to pronounce single syllable nonwords (but only after extensive practice on words).

However, it is important to note that these latter two models continue to treat the two Ds in the word DEAD as separate orthographic forms, and as such, the learning of D in one context does not directly apply to D in the other. Although the dispersion problem was reduced to the point that the models could learn to read single syllable words and nonwords, the problem becomes more serious when dealing with more complex words. Consider an example taken from Davis (1999) in which a model has learned the words CAT and HOLE, and then is tested on the morphology complex word CATHOLE. The letters H-O-L-E in CATHOLE are represented in positions four, five, six, and seven, and accordingly, even though the model has been trained on the word HOLE, this training is irrelevant to naming CATHOLE given that the letters of HOLE are coded in positions one, two, three, and four– that is, the HOLE in CATHOLE is orthographically unrelated to the morphologically simple word HOLE. Perhaps (although I doubt it) the pronunciation of CATHOLE could be supported by learning other words that have letters in the corresponding positions of CATHOLE, such as FATHER, CHOCOLATE, FUNNEL, and PASSAGE, which have the letters H, 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 CATHOLE, as these latter correspondences are irrelevant to processing the meaning of HOLE coded in positions four-to-eight. And this is only one of a number of serious problems with slot-coding schemes when dealing with complex words, as discussed in detail by Davis (1999). For brief discussion, see Andrews and Davis (1999).

Note, the above problem is not a difficulty with distributed representations per se., but rather, the input coding schemes that these models currently employ (which are themselves coded in a localist format). But what needs to be emphasized is that there are currently no connectionist models with distributed lexical representations that have solved this fundamental problem, nor have any potential solutions been suggested. At the same time, input and output coding schemes that support the identification of complex word structures have already been identified and implemented in connectionist networks that rely on the localist coding schemes at both the lexical and sub-lexical levels (e.g., Davis, 1999; Grossberg 1978; Nigrin, 1993).

The solution depends upon a spatial coding scheme in which localist letter units are coded in long-term memory in a position invariant fashion, and where the positions of letters within a word are encoded by the pattern of activation across the set of active letter units. That is, letter identity and letter position are dynamically bound on the basis on the relative activation of letters, such that the same letter nodes are used regardless of position. In these models, sequential letters are coded with decreasing activation values, and the input layer is designed so that there is a constant ratio between the activities of successive items, what Grossberg (1978) calls the invariance principle. So, CATHOLE might be represented as in Figure 1. One consequence of the invariance principle is that the pattern of activation for HOLE in CATHOLE and the morphologically simple word HOLE are the same, allowing direct

access to HOLE regardless of context. Note, repeated letters do not pose a problem for this scheme (Bradski, Carpenter, & Grossberg, 1994; Davis, 1999).

In addition to supporting the identification and naming of more complex word (and nonword) forms, this coding scheme allows networks to 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. This is actually quite difficult for many networks to achieve because superset words provide the maximum amount of excitatory input to both themselves and their subset patterns; for example, SELF activates all the letters for both SELF and ELF.[2] These problems are simply avoided when networks are restricted to words of a given length (e.g., McClelland & Rumelhart, 1981). Furthermore, Nigrin (1993) employed a spatial coding scheme in order to develop a model that learns to 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)–i.e., the model could correctly parse the input pattern THEDOGRAN. So, rather than the standard approach of recognizing words embedded in a larger pattern by first parsing words at boundaries (for example, attempting to parse spoken words based on stress, phonological redundancy, etc.), this model could parse words by first recognizing them. Davis (1999) applies a similar approach to Nigrin (1993) but focuses on visual word identification.

It is also worth mentioning that a spatial coding scheme constitutes a working memory system, consistent with the strong connection between working memory and the ability to learn new representations (e.g., Baddeley, Gathercole, & Papagno, 1998). That is, the coding scheme allows networks to represent order amongst a limited set of simultaneously active items (both at the letter and word levels), functions that are currently beyond the capacities of networks with distributed representations, as discussed later. Accordingly, this framework is a good example of a small set of general principles satisfying a wide range of cognitive demands.

In sum, connectionist networks with distributed coding schemes cannot yet accommodate the identification or naming of multi-syllabic or morphologically complex words, 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 functions.

Generalizing within regular or quasi-regular domains.

As noted above, the majority of simulation studies that have used distributed coding schemes have focused on the mapping between orthographic and phonological representations, a case in which the mappings are largely systematic and thus well suited for distributed coding. Under these conditions, distributed codes at the lexical level can support generalization, for example, supporting the pronunciation of simple orthographic strings that have not been presented previously, such as BLAP. Of course, these models can also learn input-output mappings that are not entirely systematic given their success in reading irregular words. But it is the consistent mappings that allow the distributed representations to generalize.

Although the ability to generalize is considered one of the key properties of distributed representations, there may be serious limitations to this capacity. Indeed, Hummel and Holyoak (1997; Holyoak & Hummel, in press) argue that networks that reject symbols (that would include all current networks with distributed representations) are, in principle, unable to generalize in a systematic fashion. For example, learning the relations "John is older than Mary" and "Mary is older than Sue" does not allow these networks to generalize to "John is older than Sue", because John, Mary, and Sue are all unrelated concepts in the different contexts. What is needed according to these authors is a way to ensure that John, Mary, and Sue are represented in the same way (by the same units) regardless of context. Note, this is the same problem faced by slot coding schemes in word identification, in which the letter A in first position is different from the letter A in second position. The solution proposed by Hummel and Holyoak (1997) is qualitatively the same as that proposed for letters; namely, it is necessary to represent the concepts John, Mary, and Sue independent of context, and temporarily bind these units to the role they are playing in a particular context. Indeed, these authors have implemented a model that can dynamically bind concepts such as John, Mary, and Sue to functions such as older (x, y), allowing the model to generalize in a systematic fashion. These authors also provide a mathematical analysis showing that these functions cannot be accomplished with various other approaches that have been tried, such as conjunctive coding schemes (see Hummel & Holyoak, 1997, for details).

A striking example of the limitations of networks that reject rules—such as networks that learn distributed representations via back-propagation—is their difficulty (or perhaps inability) to learn the identify function (Marcus, 1998). Marcus trained a variety of standard network models to map the first node in an input layer to the first node in an output layer, the second input node to the second output unit, etc, continuing until input node n-1 was mapped to output node n-1. This is intended to simulate the situation in which a person is trained to learn the mapping: one → one, two → two, three → three, etc. However, despite extensive training on all the input-output mappings between units 1 and n-1 (regardless of the size of n), these models failed to map the input n to the output n. This is like a human failing to respond 10 given the input 10 after learning the equivalencies from 1 to 9. The reason these models did not learn outside its training space is that they only learned to associate specific inputs with specific outputs. By contrast, humans can generalize outside the training space (even generalizing outside the number space, mapping duck to duck), because, according to Marcus, humans learn an abstract rule, namely, the identity function input(x) → output(x). Interestingly, the neural network model by Hummel and Holyoak (1997) that can learn abstract rules (including the identify function) depends upon localist coding schemes for its functioning.

There is a complication to this story, however. In a behavioral study, Marcus et al. (1999) familiarized 7-month old infants with sequences of syllables generated by an artificial grammar. The infants were then able to distinguish between sequences generated by this grammar, even when the sequences in the familiarization and test phases employed different syllables. That is, the infants showed learning outside the training set. Marcus took these data to support the existence of rules in human learning, and claimed that models that learn with back-propagation could not support this performance. However, in a response to this article, Altmann and Dienes (1999) noted that a model developed by Dienes, Altman & Gao (1999) could also generalize

beyond the training set, and indeed, the authors were able to simulate the data reported by Marcus et al. (1999) amongst other findings, despite the fact that the model included distributed representations (and no rules).

For present purposes, two points should be noted. First, the network developed by Dienes et al. (1999) requires training on all the test items for at least one iteration before any transfer is obtained from the familiarization to test phase. That is, when learning an implicit grammar, the model learns more quickly at test when the implicit grammar in the familiarization and test phases are the same, consistent with human performance (e.g., Altmann, Dienes, & Goode, 1995). However, this is a different situation to the case in which a person generalizes to items that have not been previously presented. So for instance, in the case of the identify function, a person can generate 10 in response to 10, without any previous experience with 10. The model of Hummel and Holyoak (1997) is able to generalize to completely novel input forms, although again, it relies on rules and localist representations. Second, when learning an implicit grammar, the model of Dienes et al. (1999) sometimes required more training than people on the test set before showing any transfer from the familiarization phase. The authors suggested that this may be due to the fact that standard back-propagation is poorly suited for learning complex categorizations because it learns distributed representations, and propose that the problem may be addressed by including learning rules that develop localist representations (cf. Kruschke, 1993).

So, at the very least, models with localist representations can generalize as well as models with distributed representations (cf. Page, 2000)– and if Hummel and Holyoak (1997), Marcus (1999), and others (e.g., Fodor & Pylyshyn, 1988) are correct, they have the potential to support a much more powerful form of generalization.

<u>The role of distributed representations in arbitrary input-output mappings: None.</u>

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 key functional attributes of distributed coding schemes, namely, their ability to generalize, is not an asset in these conditions. This undermines much of the adaptive value of distributed coding in the "triangle" model of reading advocated by Seidenberg, Plaut and colleagues given the majority of mappings that need to be learned are arbitrary, as the orthographic-semantic and phonological-semantic correspondences are largely unsystematic, and only orthographic-phonological codes are systematically related.

Furthermore, although unsystematic correspondence outnumber systematic correspondences in the triangle model by a ratio of 2:1, the model underestimates the number of arbitrary mappings that must be learned. For example, syntactic features of words are also often arbitrarily related to their orthographic, phonological and semantic attributes (cf. Bowers, Vigliocco, Stadthagen-Gonzalez, & Vinson, 1998).

The complexity of these mappings is important to note because arbitrary mappings are more difficult to learn using distributed compared to localist coding schemes (requiring more training), and incorrect generalizations tend to be produced. What often facilitates learning in these situations is the inclusion of more of hidden units (e.g., McRae, de Sa, & Seidenberg, 1997; Plaut, 1997). For example, in trying to model lexical decision in a distributed representations, 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 ART that has a unit for every word, learning arbitrary mappings is easier still. The important point to note is that the inclusion of more hidden units allows a network to develop more localist representations -- the form of representation that is rejected by these theorists.

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 work by Plaut and Shallice (1993) who 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 the semantic pattern log to the input dog) as well as mixed visual-and-semantic errors at a rate greater than would be expected by chance (e.g. settling into hog in response to the input dog). This pattern of errors was found to be quite general, extending to various network architectures with distributed representations that included recurrent connections (Plaut & Shallice, 1993). The ability to accommodate this complex pattern of results with a single lesion is an improvement over previous accounts that needed to assume multiple lesion sites. This success was taken to support the conclusion that distributed coding schemes do indeed support orthographic-semantic mappings.

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 recurrent 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 produce sheet), various sorts of semantic errors (e.g., output cat instead of dog), and importantly mixed errors (e.g., output rat to cat) that were more common than would be expected by 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. (Plaut, 1999, also

notes this point).

In sum, it is unclear what adaptive value distributed coding schemes play when mapping between arbitrary domains, and the arbitrary mappings outnumber the systematic mappings in the triangle model of reading advocated by Seidenberg and colleagues. Furthermore, there is no evidence that the distributed codes map between arbitrary domains, despite initial claims.

Language processing beyond single words.

Of course, even if distributed representations play no functional role in the mapping between two unrelated domains, distributed coding schemes may nevertheless be used in these situations. There may be some more basic constraints in how neurons learn that results in the formation of distributed representations, much like distributed representations develop in connectionist systems that learn arbitrary mappings with back-propagation. The more critical point is whether models with distributed coding schemes can support various key functions required for reading words and sentences (and cognition more generally). As noted above, current connectionist networks with distributed coding schemes cannot support the processing of complex word forms presented in isolation, and the problems become more severe when considering how to represent multiple pieces of information simultaneously, or representing order amongst a set of items.

To illustrate the first problem, perhaps it is worth describing the limitations of some specific connectionist models that only include distributed knowledge. First, consider a set of connectionist model that represent semantic knowledge using distributed coding schemes (Becker, Moscovitch, Behrmann, & Joordens, 1997; Borowky & Masson, 1996; Joordens & Becker, 1997; Masson, 1995; 1998; McRae et al., 1997; Plaut, 1996; Plaut & Booth, in press). 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 doctor facilitates the encoding of the related target nurse more than an unrelated prime table (although see Dalrymple-Alford & Marmurek, 1999, for some complications). This occurs because the pattern of activation generated by the a related prime is similar to the pattern of activation generated by the target, making it easier to arrive at the target state compared to a condition in which the prime is unrelated to the target.

A key problem with all these accounts, however, is that only a single concept can be represented at a time. A pattern of activation across all the units defines a single concept, and overlapping two patterns on the same set of units results in nonsense because there is no way to determine which features belong to which concept. By contrast, it seems unlikely that thoughts are composed as a series of single concepts represented one at a time in sequence, any more than sentences can be coded as a linear string of words (e.g. Vigliocco & Nicol, 1998). Rather our thoughts and language are supported by systems that encode the relations between multiple items that are co-active in parallel. Without this capacity, we can't entertain novel thoughts, such as "The sponge is larger than the tadpole". A behavioral manifestation of co-active representations within conceptual and language systems can be found in a

classic form of speech error; namely, lexical exchanges, such as: Writing a letter to my mother --> 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. It is unclear whether these errors can be explained without assuming the co-activation of multiple conceptual and linguistic representations (cf. Dell, Burger & Svec, 1997). It certainly has not been done thus far. So, although connectionist networks with distributed representations can accommodate various semantic priming data, they cannot represent meaning.

Or consider a second example, but in phonological processing. As noted above, connectionist models with distributed representations can map and generalize between orthographic and phonological knowledge given the quasi-regular relation between these two domains. But, phonological knowledge supports various functions above and beyond single-word (and pseudoword) pronunciation, such as storing multiple pieces of information in a short-term phonological store (Baddeley, 1986). Again, this is problematic, because these connectionist models of reading provide no means with which to encode more than one word (or pseudoword) at a time.

Given the above considerations, it is interesting to note that Kawamoto and colleagues (Kawamoto, 1993; Kawamoto, Farrar, & Kello, 1994) have developed a distributed theory of semantics that they claimed represents multiple semantic patterns in parallel. In particular, the authors developed a connectionist model to account for semantic ambiguity resolution. That is, there is a variety of evidence that homographic words with single spellings (or pronunciation) but multiple meanings (e.g., bank associated with money and river) activate all meanings for a brief moment, followed quickly by a process of settling down onto a single appropriate meaning (e.g., Simpson, 1984). Evidence comes from cross-modal priming studies, in which homograph primes are presented visually, and semantic priming is assessed for auditorily presented targets related to both meanings of the homograph. Under many circumstances, semantic priming is obtained to both targets, and in some studies, even when the context should rule out one meaning. For example, the prime word bugs in the sentence “Because he was afraid of electronic surveillance, the spy carefully searched the room for bugs” led to semantic priming for the auditory targets microphone and insect when the targets were presented immediately after the word bugs. But following a delay of about 750-1000 ms, semantic priming is restricted to the appropriate target microphone (Swinney, 1979).

To simulate this pattern, Kawamoto trained a network on a set of arbitrary mappings, with a given input pattern (an orthographic representation) associated with a given output pattern (a semantic representations) on some learning trials, the same input pattern associated with a different (unrelated) output pattern on other learning trials. When the input was later presented to the network, Kawamoto claimed that the network temporally represented both semantic patterns, as semantic priming could be obtained to both meanings of the word during the early stages of processing the input pattern. And consistent with the behavioral data, semantic priming was restricted to one meaning after the input was processed more extensively.

However, there are problems. First, as noted by Besner and Joordens (1995), the network actually does a very poor job of representing semantic information in

many instances. When the two alternative meanings have been learned to a similar extent, the network tended to settle into a nonsense blend of the two patterns. But even in cases in which the network actually does settle into the dominant semantic pattern, and in which semantic priming can be observed for both meanings at the early stages of processing the words, it is incorrect to conclude that both patterns were simultaneously active. Instead, the network accomplished semantic priming in the same way as the previous networks; that is, by varying the difficulty of moving from one activation pattern to another. More specifically, after the input is presented, the activation pattern for the more frequent semantic representation slowly develops, with the activation values of those semantic units with consistent values in the two patterns achieving their final values the most quickly, and those units with inconsistent values developing more slowly, being "undecided" in the early stages of processing. Because the consistent units settle more quickly into their final states, the network is actually closer in semantic space to the secondary word meaning than a completely unrelated word at the beginning of this process (in this network, each node took the value of either +1 or -1, and as a consequence, approximately half of the units had consistent values in the unrelated patterns). Given the closer activation pattern to the secondary meaning, semantic priming is obtained, as described above. But to say that the network represents the alternative word for a short interval of time is misleading, as fully half of its features are never activated, and indeed, from the very earliest stages of activation, the nodes with inconsistent values slowly settle into the value consistent with the dominant meaning. Of course, the difficulty in representing multiple forms is exactly the same as the models mentioned above; namely, the activation of one meaning is inconsistent with all other meanings, and attempting to co-activate two items simultaneously results in nonsense.

In response to these problems, one might try to argue that our brains don't actually represent multiple pieces of information in parallel, but instead learn to represent order amongst a set of items that are activated in sequence. So for example, rather than representing information in the phonological loop as a collection of co-activated items, it could be argued that items that are activated sequentially. And indeed, connectionist models with distributed coding schemes have been developed that can represent ordered information; for instance, recurrent networks developed by Elman (1990). It is interesting to note, however, that recurrent networks are 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). But more importantly for present purposes, the ability to represent sequential order in these recurrent networks was the product of slow learning mediated by changes in the connection weights between units, producing a long-term representation of ordered information in the training set. What is needed to model phonological short-term memory is an ability to temporarily represent a series of items encoded in real-time.

A solution of this sort was recently proposed by Brown, Preece, & Hulme (2000), who developed a model of short-term memory in which an internally generated context representations are associated with the to-be-remembered items. Briefly, the authors included an oscillator within their model that consists in a pattern

of activation over a collection of units that changes over time in a deterministic fashion. This oscillator is taken to represent an internal state of the model (mind) that is unaffected by the outside environment. The to-be-remembered items are also coded in a distributed fashion, and each item in turn is associated to the current (different) state of activation within the oscillator, based on Hebbian learning. At the time of retrieval, the initial state of the oscillator is reinstated, which in turn retrieves the first item in the study list based on the learned association. Because the oscillator is deterministic, it repeats the same sequence of activations from before, allowing the internal context to retrieve the associated items studied earlier. The authors use the analogy of a clock, in which the hands of the clock represent the internal context that changes over time in a deterministic manner, and just as a clock repeats itself when it is set back in time, so does the oscillator. Based on this system, the authors are able to account for a wide range of phenomena in the short-term memory literature (for related account, see Lewandowsky, 1999).

Clearly, the success of the model demonstrates that sequential information can be coded and retrieved using connectionist networks that represent information in a distributed fashion. But note, all the work in representing serial order is performed by a deterministic oscillator that is built into the model and that is unaffected by the external environment and that can be reset to repeat itself. Although oscillating systems have been observed in the brain (e.g., Gray, Konig, Engel, & Singer, 1989), there is no evidence that they have this property. And again, it is important to emphasize that this oscillator model cannot represent two things at the same time. Although this capacity may be unnecessary in order to account for short-term memory phenomena, it remains to be seen how various semantic and language skills can be explained without co-active representations.

To summarize this section thus far, current networks with distributed representations cannot represent multiple items simultaneously, and although order can be represented on-line, this has only been accomplished by building a great deal of purpose-built structure into the network—exactly what authors from the PDP camp claim to avoid.

In contrast with these limitations, it is important to emphasize that there are connectionist models of semantic and phonological knowledge that do represent multiple pieces of information simultaneously and on-line, but they all rely localist representations in order to achieve this function. For example, in the case of semantic knowledge, Hummel and Holyoak (1997) developed a model capable of representing complex propositions by representing multiple semantic representations in parallel, employing a complex interplay between distributed and localist codes, where localist representations function to bind together the semantic features that belong to one representation or another (thus avoiding the nonsense blend that results in other networks that avoid localist representations). In the case of phonological knowledge, Page and Norris (1998) and Burgess and Hitch (1999) have developed localist connectionist systems in which order information is represented in terms of the relative activation of co-active localist codes—that is the spatial coding schemes used to identify complex word forms (e.g., Davis, 1999)—and like the Brown et al. (2000) model, these models can account for a wide range of short-term memory phenomena (also see Grossberg & Stone, 1986; Nigrin, 1993).

Similarly, models with localist representations can learn one to many mappings, such as mapping between the orthographic form bank to the semantic representations money-bank and river-bank; in particular, networks within the ART family (cf. Bartfai, G., 1996; Carpenter & Markuzon, 1998). One reason ART models can learn these mappings is that they have solved the so-called stability-plasticity problem (Grossberg, 1978; also known as catastrophic interference, McCloskey & Cohen, 1989; Ratcliff, 1990) such that new learning does not erase past knowledge. Briefly, these mappings can be learned in ART networks that include a teacher, so-called ARTMAP networks. Imagine the case in which a specific pattern of activation in an input layer learns to activate a specific node in a second layer (call it X). Now, if the same input-output mapping is determined to be an error on a later trial, based on the feedback from the teacher, then the active node X is automatically turned off, and a new node is activated in the second layer (node Y). As a consequence of activating Y, connection weights are changed between the two layers so that Y is activated based on this same input. The key point, however, is that the connection weights between the input units and X are unchanged because learning only occurs between active items in the two layers, and X was turned off before any learning took place. As a consequence, when the input is now presented to the network, both X and Y are activated.[3] By contrast, because models that learn with back-propagation have not overcome the stability-plasticity problem, the attempt to learn inconsistent patterns causes the network to vacillate between the two possible outcomes, often leading to blends (or the network settles to the more frequent pattern), as noted above.

More generally, it is interesting to note that the debate concerning the relative virtues of distributed vs. localist coding schemes is largely non-existent in literature focused on sentence level processing. Rather, the debate has focused on how many localist lexical representations are required. For example, in the speech production literature, Levelt and colleagues (e.g., Levelt, 1989; Levelt, Roelofs, & Meyer, 1999) have developed a network model of speech production that involves contacting two lexical representations in sequence, with a conceptual message first activating localist lexical-semantic representations (so-called lemmas) followed by the activation of localist lexical-phonological representations (so-called lexemes). Furthermore, lemmas are associated with various syntactic features coded locally, and lexemes are connected with sub-lexical localist phonological representations. By contrast, Caramazza and colleagues (e.g., Caramazza, 1997; Caramazza, & Miozzo, 1998; also see Dell & O'Seaghdha, 1994) have argued that the lemma and lexemes representations should be collapsed to a single localist representation, which in turn would be connected to localist syntactic and sub-lexical phonological representations.

The widespread inclusion of localist representations can also be found in the language comprehension literature, even in cases in which the authors are normally committed to distributed representations (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994). Presumably, the reason that these models have tended to included localist knowledge is that the authors are confronting head on the problem of encoding sequences of words on-line, and localist representations seem to serve an important function in these situations.

So in sum, connectionist models that learn distributed representations via back-propagation and related procedures cannot represent multiple-items in parallel, and their ability to represent order on-line depends upon built in structures, such as

deterministic oscillators. By contrast, networks that learn localist lexical representations using spatial coding schemes can represent multiple items in parallel, and this capacity supports their ability to represent order on-line.

Overall summary

The key theoretical contribution of connectionist models is not that they provide a means to accommodate various cognitive phenomena while rejecting all lexical localist representations, 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 et al., 1986). And by building models from basic principles, the 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 important, current models that include distributed coding fail to support a wide variety of cognitive functions. The greatest success of distributed coding schemes have been found in domains that involve quasi-systematic input-output mappings, in particular, the task of naming words and nonwords. But even within this domain, current models are restricted to processing single syllable items, and serious empirical and computational challenges need to be met before they adequately account for human performance on these simple forms, let alone human performance on more complex structures, such as novel compound words. Outside this domain, when the input-output mappings are largely arbitrary (which is the majority of cases in the triangle model of reading supported by Seidenberg, McClelland, Plaut, and colleagues), it is unclear what function distributed representations play. But most problematic, models with distributed representations have not been applied to more complex cognitive phenomena that are fundamental to reading and cognition more generally, such as representing more than one word at a 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 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, and they support all the functions of distributed coding schemes, as well as more complex computations 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 order amongst a set of items in short-term memory (e.g., Bradski, Carpenter, & Grossberg, 1994; Page & Norris, 1998), and can represent multiple items simultaneously (e.g., Cohen & Grossberg, 1987; Nigrin, 1993). 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 his ART architectures around a small set of key functional demands, including (a) the ability to process information in noise, such 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, (c) the ability to learn in real time with or without a teacher, (d) the ability to identify and 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 satisfied using the ART and related frameworks, and have been applied to a wide range of cognitive phenomena, as noted above. Before localist coding schemes are rejected, models with distributed coding schemes should at least match this performance.

Concluding comment

Despite the computational power and neural plausibility of connectionist network that learn localist representations, in particular the ART models of Grossberg and colleagues, this work is almost entirely ignored in the psychological literature. The level to which this work is ignored is quite striking. For instance, in developing a possible solution to the stability-plasticity problem, McClelland, McNaughton, and O'Reilly (1995) list many past attempts to solve the problem, and conclude that all these approaches reduce catastrophic interference at the cost of their ability to generalize from past experience. However, they fail to mention ART networks which were specifically designed to solve this problem. Indeed, McClelland et al. (1995) mistakenly credit McCloskey and Cohen (1989) and (Ratcliff, 1990) for independently identifying the stability-plasticity problem (what they labeled catastrophic interference), whereas Grossberg identified and provided a formal solution to this problem much earlier (Grossberg, 1976). Ratcliff (1990) does credit Grossberg (1987): "Grossberg has forcefully argued that many of the recent network models are unstable under temporally changing training environments, and the problem is illustrated with reference to the back-propagation learning algorithm"(p 306). However, he fails to mention that Carpenter and Grossberg (1987) implemented the ART model that solved this problem.

Although there are a few examples of researchers associated with the PDP camp quoting Grossberg (e.g., Stone, & VanOrden, 1994), the number of references to ART and related models is not far from zero. This is undoubtedly the reason why it is often assumed that distributed representations are one of the core connectionist principles. If nothing else, the field needs to more fully consider 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.[4]

Footnotes

1. Localist coding would also be implemented if the letter A was coded with 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 is uniquely involved in coding for the letter A, and another set of non-overlapping neurons uniquely involved in coding for B, etc..

2. 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 in spatial coding network, see Bowers, (1999).

3. Movellan and McClelland (1993) describe a distributed network that could in fact learn one to many mappings—although it can not represent two things at the same time. However, this network has not been tested on any psychological data, and basic questions regarding how many patterns can be learned, whether the network suffers from catastrophic interference when multiple patterns are learned, whether sequences can be learned, etc. have yet to be addressed. If this network is shown to model the same set of phenomena that are currently accommodated with localist networks, then clearly, this particular argument loses force.

4. For anyone interested in learning more about this approach, perhaps the best place to start is Grossberg (1987), who explicitly compares the ART framework with more popular learning algorithms, including back-propagation networks. Also read the excellent book by Nigrin (1993), which, in addition to making important advances, has an excellent introduction to Grossberg's work. Davis' (1999) Ph.D. thesis applies this general framework to build an impressive model of visual word identification that extends to multi-syllable and morphologically complex words, and a book is in the works.

References

Altman, G. T. M., Dienes, Z., & Goode, A. (1995). On the modality independence of implicitly learned grammatical knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 899-912.

Andrews, S., & Davis, C. (1999). Interactive activation accounts of morphological decomposition: Finding the trap in mousetrap? Brain and Language, 68, 355-361.

Baddeley (1986). Working memory. Oxford: Oxford University Press.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. Psychological Review, 105(1), 158-173.

Bartfai, G. (1996). On the match tracking anomaly of the ARTMAP Neural Network. Neural networks, 9, 295-308.

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(5), 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(2), 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(3), 432-446.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. Journal of Experimental Psychology-Learning Memory and Cognition, 22(1), 63-85.

Bowers, J. S. (1996). Different perceptual codes support priming for words and pseudowords: Was Morton right all along? Journal of Experimental Psychology-Learning Memory and Cognition, 22(6), 1336-1353.

Bowers, J. S. (1999). Grossberg and colleagues solved the hyperonym problem over a decade ago. Behavioral and Brain Sciences, 22(1), 38.

Bowers, J.S. (2000). In defense of abstractionist theories of word identification and repetition priming. Psychonomic Bulletin & Review, 7, 83-99

Bowers, J. S., & Michita, Y. (1998). An investigation into the structure and acquisition of orthographic knowledge: Evidence from cross-script Kanji-Hiragana priming. Psychonomic Bulletin & Review, 5(2), 259-264.

Bowers, J. S., Vigliocco, G., & Haan, R. (1998). Orthographic, phonological, and articulatory contributions to masked letter and word priming. Journal of Experimental Psychology-Human Perception and Performance, 24(6), 1705-1719.

Bowers, J.S., Vigliocco, G, Stadthagen-Gonzalez, H., &. Vinson, D. (1999). Distinguishing language from thought: Experimental evidence that syntax is lexically rather than conceptually represented. Psychological Science, 10, 310-315.

Bradski, G., Carpenter, G. A., & Grossberg, S. (1994). Store Working-Memory Networks For Storage and Recall Of Arbitrary Temporal Sequences. Biological Cybernetics, 71(6), 469-480.

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

Brown, H., Sharma, N. K., & Kirsner, K. (1984). The Role of Script and

Phonology In Lexical Representation. Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 36(3), 491-505.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. Psychological Review, 106(3), 551-581.

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., & Markuzon, N. (1998). ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. Neural Networks, 11, 323-336.

Caramazza, A. (1997). How many levels of processing are there in lexical access? Cognitive Neuropsychology, 14(1), 177-208.

Caramazza, A., & Miozzo, M. (1998). More is not always better: a response to Roelofs, Meyer, and Levelt. Cognition, 69(2), 231-241.

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. (1981). Disorders of reading and their implications for models of normal reading. Visible Language, 3, 245-286.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of Reading Aloud - Dual-Route and Parallel-Distributed- Processing Approaches. Psychological Review, 100(4), 589-608.

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.

Dienes, Z., Altmann, G. T. M., & Gao, S. J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. Cognitive Science, 23, 53-82.

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval In Sentence Production. Psychological Review, 93(3), 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(1), 123-147.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and Convergent Lexical Priming In Language Production - a Comment. Psychological Review, 98(4), 604-614.

Dell, G. S., & O'Seaghdha, P. G. (1994). Stages Of Lexical Access In Language Production. Cognition(SISI), 287-314.

Elman, J. L. (1988). Finding structure in time (CRL Tech. Rep. No. 8801). La Jolla, CA: UCSD.

Elman, J. L. (1990). Finding Structure In Time. Cognitive Science, 14(2), 179-211.

Feldman, L. B., & Moskovljevic, J. (1987). Repetition Priming Is Not Purely Episodic In Origin. Journal of Experimental Psychology-Learning Memory and Cognition, 13(4), 573-581.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive

Architecture - a Critical Analysis. Cognition, 28(1-2), 3-71.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read- out model. Psychological Review, 103(3), 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.

Gray, C.M., Konig, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar syncronization which reflect global stimulus properties. Nature(338), 334-337.

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. (1982). Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control. Boston: Reidel Press.

Grossberg, S. (1985). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), Pattern Rocognition by Humans and Machines (Vol. 1). New York: Academic Press.

Grossberg, S. (1987). Competitive Learning - From Interactive Activation to Adaptive Resonance. Cognitive Science, 11(1), 23-63.

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

Grossberg, S., & Stone, G. (1986). Neural Dynamics of Attention Switching and Temporal-Order Information In Short-Term-Memory. Memory & Cognition, 14(6), 451-468.

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

Hinton, G. E., & Shallice, T. (1991). Leasoning an Attractor Network - Investigations of Acquired Dyslexia. Psychological Review, 98(1), 74-95.

Holyoak, K. J., & Hummel, J. E. (in press). The Proper Treatment of Symbols in a Connectionist Architecture. In E. Dietrich & A. Markman (Eds.), Cognitive dynamics: Conceptual change in humans and machines . Cambridge, MA: MIT Press.

Hummel, J. E., & Biederman, I. (1992). Dynamic Binding In a Neural Network For Shape-Recognition. Psychological Review, 99(3), 480-517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104(3), 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(5), 1083-1105.

Kawamoto, A. H. (1993). Nonlinear Dynamics In the Resolution of Lexical Ambiguity - a Parallel Distributed-Processing Account. Journal of Memory and Language, 32(4), 474-516.

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(6), 1233-1247.

Kruschke, J. K. (1993). Human category lerning: Implications for backpropagation models. Connection Science, 5, 3-36.

Levelt, W. J. M. (1989). Speaking: From Intetion to Articulation. Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). Multiple perspectives on word production. Behavioral and Brain Sciences, 22(1), 61-75.

Lewandowsky, S. (1999). Redintegration and response suppression in serial recall: A dynamic network model. International Journal of Psychology, 34, 434-446.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical Nature Of Syntactic Ambiguity Resolution. Psychological Review, 101(4), 676-703.

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

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science, 283(5398), 77-80.

Masson, M. E. J. (1995). A Distributed-Memory Model of Semantic Priming. Journal of Experimental Psychology-Learning Memory and Cognition, 21(1), 3-23.

McClelland, J.L. (1976). Preliminary letter identification in the perception of words and nonwords. Journal of Experimental psychology: Human Perception and Performance, 3, 80-91.

McClelland, J. L., McNaughton, B. L., & Oreilly, R. C. (1995). Why There Are Complementary Learning-Systems In the Hippocampus and Neocortex - Insights From the Successes and Failures Of Connectionist Models Of Learning and Memory. Psychological Review, 102, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects In Letter Perception .1. an Account of Basic Findings. Psychological Review, 88(5), 375-407.

McClelland, J., Rumelhart, D. E., & the PDP Research Group (1986). Parallel Distributed Processing: Psychological and Biological Models. (Vol. 2). Cambridge, MA: MIT 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(2), 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.

Movellan, J. R., & McClelland, J. L. (1993). Learning Continuous Probability-Distributions With Symmetrical Diffusion Networks. Cognitive Science, 17(4), 463-496.

Nigrin, A. (1993). Neural Networks for Pattern Recognition. Cambridge, MA: MIT Press.

Page (in press). Connectionist modeling in psychology: A localist manifesto. Behavioral and Brain Sciences.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. Psychological Review, 105(4), 761-781.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. Brain and Language, 52(1), 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(5-6), 765-805.

Plaut, D. C. (1999). Computational modeling of word reading, acquired dyslexia, and remediation. In R. M. Klein & P. A. McMullen (Eds.), Converging Methods for Understanding Reading and Dyslexia . London: MIT Press.

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

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(1), 56-115.

Plaut, D. C., & Shallice, T. (1993). Deep Dyslexia - a Case-Study of Connectionist Neuropsychology. Cognitive Neuropsychology, 10(5), 377-500.

Rastle, K., & Coltheart, M. (1997). Whammy and double whammy: Inhibitory effects in reading aloud. Paper presented to the thirty eighth annual meeting of the Psychonomic Society, Philadelphia.

Rastle, K., & Coltheart, M. (1999). Lexical and nonlexical phonological priming in reading aloud. Journal Of Experimental Psychology-Human Perception and Performance, 25(2), 461-481.

Ratcliff, R. (1990). Connectionist Models Of Recognition Memory - Constraints Imposed By Learning and Forgetting Functions. Psychological Review, 97(2), 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., & Zipser, D. (1985). Feature Discovery By Competitive Learning. Cognitive Science, 9(1), 75-112.

Seidenberg, M. S. (1993a). Connectionist Models and Cognitive Theory. Psychological Science, 4(4), 228-235.

Seidenberg, M. S. (1993b). A Connectionist Modeling Approach to Word Recognition and Dyslexia. Psychological Science, 4(5), 299-304.

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review, 96(4), 523-568.

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.

Simpson, G. B. (1984). Lexical Ambiguity and Its Role In Models Of Word Recognition. Psychological Bulletin, 96(2), 316-340.

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

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.

Swinney, D. (1979). Lexical access during sentence comprehension: Reconsideration of some context effects. Journal of Verbal Learning and Verbal Behavior, 18, 645-659.

Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? Cognition, 68, B13-B29.