Parallel Activation of Distributed Concepts: Who put the P in the PDP?

M. Gareth Gaskell
Centre for Speech and Language,
Psychology Department, Birkbeck College,
Malet Street, London WC1E 7HX, England
g.gaskell@psyc.bbk.ac.uk

Abstract

An investigation of the capacity of distributed systems to represent patterns of activation in parallel is presented. Connectionist models of lexical ambiguity have captured this capacity by activating the arithmetic mean of the vectors representing the relevant meanings to form a lexical blend. However, a more extreme test of this system occurs in a distributed model of lexical access in speech perception, which may require a lexical blend to represent transiently the meanings of hundreds of words. I show that there is a strict limit on the number of distributed patterns that can be represented effectively by a lexical blend. This limit is dependent to some extent on the structure and content of the distributed space, which in the case of lexical access corresponds to structure and content of the mental lexicon. This limitation implies that distributed models cannot be simple re-implementations of parallel localist models and offers a valuable opportunity to distinguish experimentally between localist and distributed models of cognitive processes.

Introduction

One of the cornerstones of the connectionist enterprise is the representation of information in a distributed fashion: Each pattern is represented over many processing units and each processing unit forms part of many patterns. This contrasts directly with localist systems, in which each concept is represented by the activation of a single word. Localist models have been valuable in modeling perceptual processes in which the degree of match between sensory input and a set of possible candidates for identification can be represented in terms of a set of activation values; each candidate having a separate activation (e.g., Morton, 1969). The essential point about this type of system is that there is no limit to the number of candidates that can be activated in parallel, since each is independently represented.

Activation of multiple candidates in distributed networks has been achieved by averaging or "blending" the relevant vectors to form a pattern similar to all its constituents. Multiple candidates can be said to be activated to the extent that they are near to the blend in vector space (e.g., Kawamoto, 1993). However, it is not clear whether this approach represents a literal re-implementation of localist activations or whether it is merely an approximation to the localist systems, with inherent limitations.

In this article, I put the distributed blending approach through its paces, examining a variety of lexical representations. I tackle this problem from an abstract perspective, rejecting actual network simulations in favor of simple mathematical and statistical analyses of vector spaces. This allows a wide range of relevant parameters to be explored without restricting the scope of the analysis to one particular network architecture or learning algorithm.

These analyses are discussed with reference to localist and distributed models of human speech perception. Speech perception provides an important test-bed for questions of parallel activation, partly because the field has been dominated by models in which word candidates are represented in a localist fashion (e.g., Marslen-Wilson, 1987; McClelland & Elman, 1986). More importantly, however, the temporal nature of speech allows us to examine parallel activation during the time-course of perception of words (e.g., Zwitserlood, 1989). This creates the potential for experimentally distinguishing between localist and distributed models of cognitive processing.

A Distributed Model of Speech Perception

Parallel models of speech perception such as Cohort (Marslen-Wilson, 1987) and TRACE (McClelland & Elman, 1986) assume that as a word is heard, many word candidates are assessed simultaneously. The Cohort model goes further, arguing that as these word candidates are evaluated their meanings also become activated. Experimental evidence for this behavior comes from priming studies (e.g., Zwitserlood, 1989), in which an ambiguous word onset (e.g., /kæpt/) facilitates the recognition of targets related in meaning to more than one possible continuation of the stimulus (e.g., *ship* related to *captain*, *prison* related to *captive*). However, the extent to which parallel activation occurs (i.e., whether it extends to large cohorts) remains unknown.

Cohort and TRACE are essentially localist models, in which the goodness of fit between each word candidate and the incoming speech is represented by a separate activation value. Gaskell & Marslen-Wilson (1995) examined the effects of implementing the lexical access process for speech in a distributed learning system. They trained a simple recurrent network to learn the mapping from a stream of phonetic features (segmented into phoneme-like

units) onto distributed representations encompassing the meaning and phonological form of words. Lexical access is interpreted in terms of movement through a multi-dimensional space, with word representations being fixed points in this space (see Figure 1). The output of the network plots the course of this movement: As speech information gradually enters the network, the activation of matching words is reflected by constructing a blend of their distributed representations. When the onset of a word is presented at the input, the network outputs a blend of the representations of all the words containing that onset. As more speech comes in, this blend can be refined to represent the reduced set of words that still match the speech input. This refinement continues until the number of words matching the input reduces to 1. At this point (the uniqueness point) the network can isolate the full distributed representation of the remaining word: It has reached an endpoint in the lexical space.

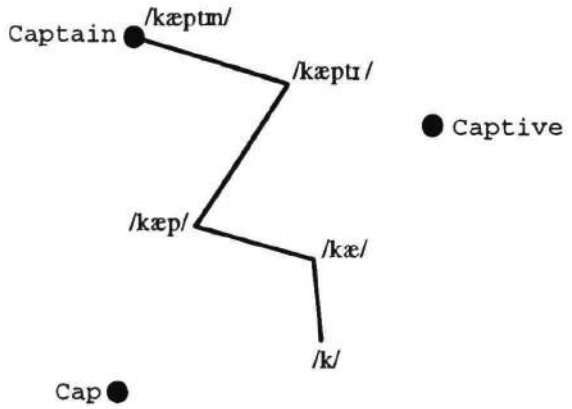

Figure 1. Lexical access as a trajectory through lexical space. The dots mark word representations and the line marks the path of the network output vector as speech is processed.

Lexical Distance and Activation

Localist models of auditory lexical access use the activation metaphor to indicate the status of recognition process—the degree of match between each word and the incoming speech is reflected in the word's activation value. In the distributed model, this activation is encoded implicitly by the position of the output vector in lexical space—the degree to which any word's lexical representation has been retrieved depends on the proximity of the representation to the output of the network. This proximity value is highly dependent on the number of words that must be activated. If the uniqueness point of a word has been reached, the network merely has to reproduce the lexical representation of that word and so the distance between the output of the network and the representation of that word is likely to be small. This corresponds to a high degree of activation for the word in a localist model. If (as in the *captain/captive* case) the input is temporarily consistent with two words, the network can at best output a value half way between the corresponding points in lexical space. When more words are part of a lexical blend, the distance between the blend and the component words is greater (and thus in localist terms, their activations are smaller).

Figure 2 (bold line) illustrates this pattern using a randomly defined lexical space with 200 binary dimensions. Each word is represented by a vector, with each element of the vector having a 50% chance of being on or off. Sets of target patterns were randomly selected and a blend vector was calculated by taking the mean over all target values for each element. The root-mean squared (RMS) distance from this blend vector was then calculated for all the target vectors. Each point in Figure 2 is based on the mean of 64 values. As the number of target patterns increases, their distance from the blend also increases. Thus, word activation as modeled by proximity is highly dependent on the number of candidates remaining active.

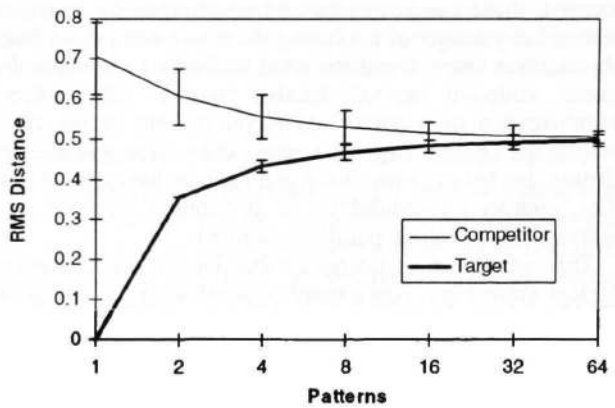

Figure 2. Mean, maximum and minimum distances from blends of targets to target and competitor populations.

It is important to relate the distance of these lexical blends from target representations to the overall population of distances. To be an effective representation of its components a blend should not only be close to the components, it should also be relatively distant from other words. It follows that representational effectiveness can be evaluated in signal detection terms: a blend is an effective representation of its components (the targets) if the target and competitor populations can be separated on the basis of lexical distance alone. Figure 2 also plots the distances from the lexical blends to a set of vectors representing unrelated words in the network's mental lexicon: 3000 randomly chosen vectors with the same properties as the target vectors.

When the number of target patterns is small, lexical blends are much closer to those target vectors than to any of the competitor words. For example, when the lexical blend is based on two target patterns, the RMS distance between those patterns and the blend is 0.36. This is comfortably closer than the nearest competitor, which is 0.53 from the blend. However, as the number of target patterns increases, the signal begins to merge with the

noise and the blends become less informative. It soon becomes impossible to work out which of the words the blend is intended to represent on the basis of proximity. It seems that modeling parallel activation in this way imposes a limit on the number of words that can be usefully activated. If too many distributed patterns are blended together, the interference between them becomes large and there is a good chance of some spurious pattern falling closer to the blend than many of the target patterns. Hinton and Shallice (1991) show that a blend of two vectors in this type of system will always be as close to those vectors as any other vector (if not closer). However, for blends of a larger number of words this is not the case: it becomes possible (and even probable) that other vectors will fall closer to the blend than one or more of the target vectors.

It follows that a distributed system cannot implement localist activation models literally. Such models may permit many thousands of candidates to be active early in the processing of a stimulus. Because representations are localist, these candidates can be simultaneously activated without any danger of confusing the active candidates from the inactive ones. The distributed equivalent can reach the same endpoint as a localist model (the correct identification of a perceptual stimulus), but in the early stages of processing its state does not completely distinguish between matching and mismatching candidates. The number of candidates a distributed network can activate effectively in parallel is limited.

This conclusion seems reasonable, but since the assumed lexical system involves a number of arbitrary parameters it should be treated with some caution. In the following sections, I explore the extent to which using different lexical systems alter the properties of a distributed lexical access process.

Dimensionality of Lexical Space

The extent to which multiple lexical representations can be activated simultaneously in a distributed lexicon depends in part on the number of dimensions in that lexical space. In order to discriminate between the components of a lexical blend (the "active" words) and their competitors (all other words in the mental lexicon), the components must match the blend on more features than the competitors. Again, assuming competitors are randomly distributed through the lexical space, this means that each competitor will have a certain chance of matching the blend on each feature. If there is a small number of features and a large number of competitors, then the lexical space becomes crowded and there is a good chance of at least one competitor being sufficiently similar to a lexical blend to cause interference. As the number of dimensions or features rises, this likelihood diminishes and the capability of the lexical system to accommodate multiple representations increases.

This was demonstrated using randomly chosen competitor sets, again with binary dimensions and a 50% chance of each element being set to 1. We defined the separability of target and competitor populations to be the difference between the mean target distance and the minimum competitor distance. This gives a simple measure of the representational effectiveness of the blend. A high separability value implies that the two populations are separable on the basis of distance from the blend vector and indicates that the system is adequately representing the target patterns in parallel.

For each lexical space, consisting of between 50 and 800 dimensions, the separability of the target and competitor sets decreases as the number of patterns in the target increases (see Figure 3). However, as the dimensionality of the space rises, the target representations become easier to separate from the noise. This effect is most obvious in the x-axis zero-crossing points for each space, which can be thought of as a measure of the capacity of the system for simultaneous representation of distributed forms. This capacity rises from about 4 to 32 as the dimensionality of the space rise from 50 to 800. Thus, increasing the number of dimensions in the lexical space improves the capacity for activating multiple representations in parallel. It is difficult to determine where the human system lies along this continuum of dimensionality, but it may be best to think of dimensionality as a measure of richness or degrees of freedom in lexical representations. Each way of distinguishing between two words adds an extra dimension or feature to the representation and more obliquely adds to the capacity of the system to represent multiple lexical entries in parallel.

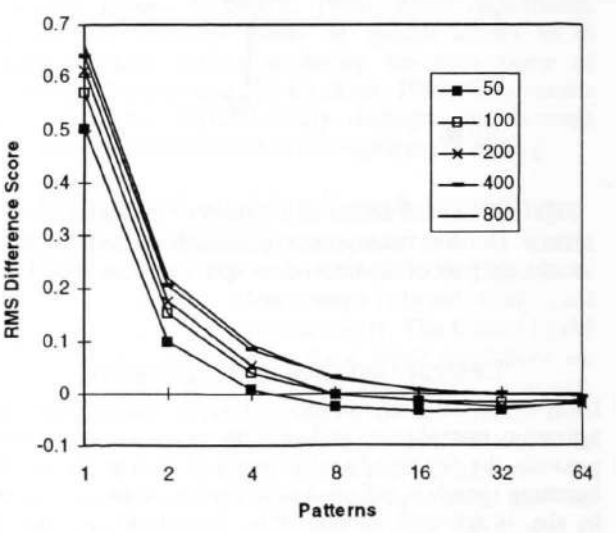

Figure 3. Effect of dimensionality on separability of target and competitor populations. The y-axis plots the minimum competitor distance minus the mean target distance.

Sparseness of Lexical Representations

Many models of cognitive functioning (e.g., Hinton & Shallice, 1991; Plaut & Shallice, 1993) have assumed that distributed lexical representations are sparse, meaning that each word's representation will involve the activation of only a small number of elements. The need for sparse representations is most obvious with binary micro-featural

representations of word meaning, where each feature is only relevant to a small minority of words. In less literal representations, sparseness may translate to a high degree of correlation between the distributed vectors representing words. This factor seems bound to affect the capacity for simultaneous activation—after all the localist position, which is ideally suited to parallel activation, occupies one end of the continuum of sparseness. The representations examined so far, in which 50% of all elements were randomly set to 1, lie at the opposite end of this continuum.

Figure 4 shows the effects of manipulating sparseness, using the separability measure defined earlier. Targets and competitor sets in a 200 dimensional space were assigned distributed representations randomly, but the probability of any element being set to 1 (p_{on}) varied from 0.05 to 0.5. Competitor set size was fixed at 3000 words. Also plotted is the same measure for a set of localist representations (local) and for a "near-localist" system of 2 elements on per word (2feat). For the localist representation, the number of competitors is limited by the number of elements in the vector, but since each competitor is equidistant from the blends, this has no effect on the results.

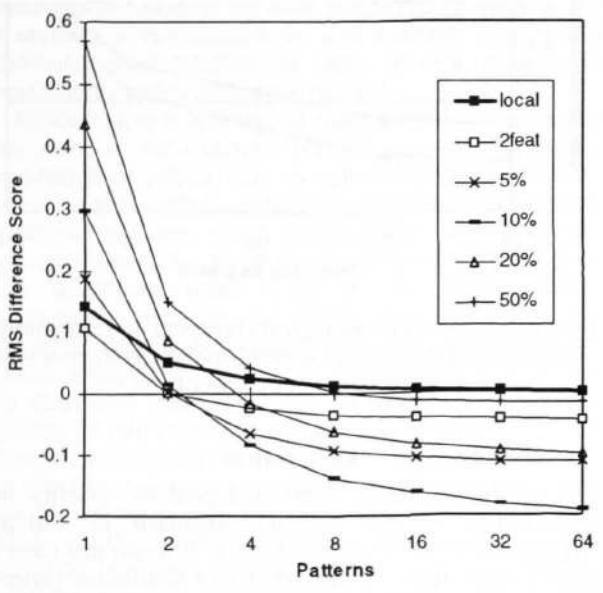

Figure 4. Effect of sparseness of representation on separability. The key gives the mean percentage of elements set to 1 per pattern. The local and 2feat curves are explained in the text.

The pattern that emerges from this manipulation is complex. This is partly because as sparseness decreases the range of possible distances in the lexical space is reduced, which has the effect of flattening the curves for the sparser representations. The most salient feature of each curve is the x-axis zero-crossing point. This marks the point at which the nearest competitor is as close to the blends as the average target and gives an indication of the point at which the signal disappears into the noise. As p_{on} reduces from 0.5 to 0.05, this zero-crossing drops from roughly 8 to 2 patterns. This implies, perhaps surprisingly, that the capacity for multiple representation drops as sparseness increases. The curve for the 2-feature representation fits in with this pattern, crossing the x-axis at roughly 2 patterns. However, the curve for the localist representation is very different: it is still (minimally) above the x-axis for a blend of 64 patterns and in fact should never cross the x-axis.

In summary, increasing sparseness in a distributed representation deepens the problem of representing words simultaneously, despite the fact that the sparser representations seem more similar to a localist representation, which is only limited by the number of elements in the vector. The sparse representations are problematic because they place a restriction on the positions in lexical space that words can occupy. This is similar to reducing the dimensionality of the space, which also reduces the capacity of the system. The localist system is crucially different: it also restricts the lexical space but it guarantees that each word is orthogonal to and equidistant from every other word. This compartmentalizes the space, meaning that a blend of any number of words will always be closer to those words than to all others.

Non-random Distribution in Lexical Space

The lexical systems examined so far have assumed that word representations are randomly distributed through lexical space. This assumption seems implausible if lexical space encodes any kind of similarity between words. Gaskell & Marslen-Wilson (1995) describe lexical access as a mapping onto word representations in a combined phonological and semantic space. Each of these types of knowledge provides structure, which shapes the lexical space and may alter the nature of the blending of representations as speech is perceived.

To address this issue, we need a distributed representation that encodes the similarity structure both of the meanings and the phonological forms of words. Lund, Burgess & Atchley (1995) have argued that similarity in meaning can be captured using co-occurrence statistics drawn from large corpora of language. This method relies on the assumption that words with similar meanings will occur in similar contexts. Although this approach is unlikely to capture the full richness of word meanings, it is a simple and convenient way to capture some aspects of semantic similarity in a distributed system.

Figure 5 compares a random lexical space to two sets of more structured representations taken from Lund et al. (1995). The structured representations are of a set of 2779 word representations (mostly of monosyllabic words). Each one is a 200 element vector with values ranging from 0 to 645. The 200 dimensions were selected from a larger matrix of co-occurrence statistics in order to capture the maximum variance between the vectors for the chosen words. The 64 target words were selected randomly from this set, with all other words acting as competitors. A second analysis used a binary form of these vectors, in which each element was set to either 1 or 0 depending on whether it was above or below the mean value across all

words. The random space also had 200 binary dimensions, with each element having a 50% chance of being set to 1.

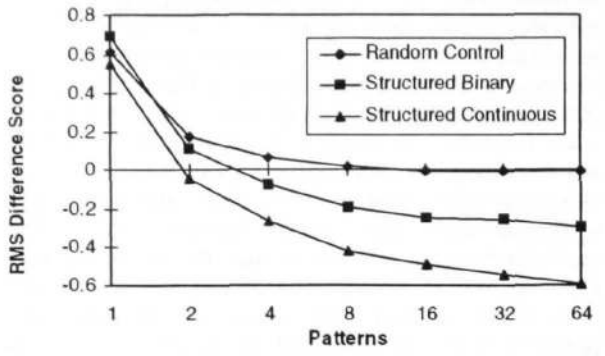

Figure 5. Effect of semantic clustering on separability. The RMS scores for the continuous space are normalized.

The separability curves show that both forms of structured vectors suffer more from the problem of blending than the random vectors. The zero crossing for the random vectors is at roughly 16 patterns, whereas for the binary structured vectors it is between 2 and 4, and for the continuous structured system it is below 2. This implies that for the latter system there is a fair chance of a blend of even 2 vectors falling closer to some other word than to the constituents of the blend.

The more realistic space has more problems distinguishing signal from noise because groups of words form tight clusters in the space. For example, representations of food words may be highly similar to each other but very different to all other representations. This means that when one of these representations is blended with the representation of an unrelated word there is a good chance of one of the other words in the cluster being as close or closer to the blend than the target. The non-binary form of this representation fares even worse because there are no restrictions on the positions word representations can occupy in the space. In particular, words may well occupy positions close to the middle of the space, which is where the blends, being arithmetic means, tend to sit.

In general, therefore, adding more realistic clustering worsens the problem of activating distributed representations simultaneously. However, there is one case in which more realistic clustering lessens this problem. For models of speech perception this is the case where lexical dimensions reflect similarities in the phonological form of words. This is because the phonological representations of words that must be activated in parallel (i.e., cohort members) will be more similar to each other than to unrelated words. Along the dimensions that encode the similarities, the blend will match the target representations exactly, but will mismatch competitors. This gives the targets a head start in terms of their overall distance to the blend in lexical space, and decreases the chances of non-cohort members falling close to the blend vector.

To illustrate this effect, target and competitor word sets were selected using cohort groupings for a word chosen randomly from the 2628 monosyllables in the Lund, Burgess and Atchley (1995) set (the word *bound*). The 223 words with /b/ as initial segment formed the target set for the first blend, with all other words treated as competitor set; the second blend used only the 25 words with onset /ba/ as targets and so on. The lexical space consisted of 52 phonological dimensions, which encoded a modified form of the Plaut, McClelland, Seidenberg & Patterson (1996) monosyllabic representation, and 52 semantic dimensions, which were random, binary and matched the phonological representations on sparseness (p_{on} = 0.08). This space was compared to a control space in which all 104 dimensions were random (see Figure 6). For both lexical spaces, the ability to separate cohort (target) from competitor sets increases further into the word, as the cohort set size decreases. However, the space incorporating phonological structure is more able to separate cohort from competitor sets at all points, reflecting the similarity between cohort members along the phonological dimensions of the lexical space.

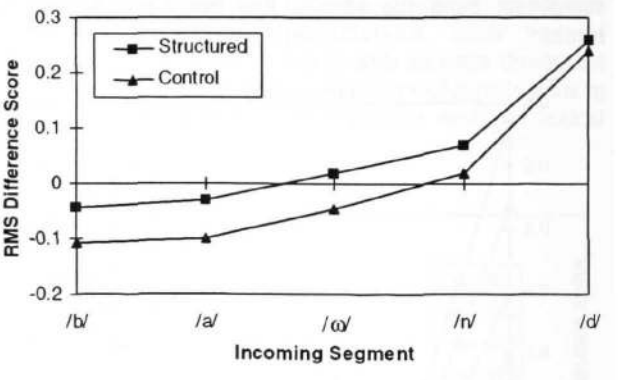

Figure 6. Effect of phonological clustering on separability. The target sets represent the word-initial cohort groups at each point in the word.

Discussion

The previous sections have attempted to quantify the effectiveness of the blending approach to multiple representations in a distributed space. It seems that there is quite a strict limit on the number of distributed patterns that can be usefully combined into a single blend. In general, combining more than a handful of representations results in an unsatisfactory blend, for which simple distance in lexical space does not properly distinguish the components of the blend from their competitors. This means that distributed networks do not simply re-implement localist, activation based systems such as the Cohort (Marslen-Wilson, 1987) or logogen (Morton, 1969) models. This conclusion, although introduced with reference to models of speech perception, may have implications for many domains of cognitive processing, such as short-term memory capacity (e.g., Miller, 1956) or conceptual combination.

Various structural factors affect the capacity for multiple representation. It correlates positively with the number of

dimensions or degrees of freedom in the lexical space. Similarly, the sparseness of lexical representations has some effect, with more sparse representations decreasing the capacity to accommodate multiple distributed representations, despite their surface similarity to localist representations. The addition of structure to the distribution of words in lexical space generally increases the problem of multiple distributed representation, because words that are closely packed together in space are difficult to discriminate on the basis of lexical distance alone. This problem becomes more acute when the dimensions of lexical space are continuous rather than binary. The one case in which the addition of structure does help is when the target patterns are all similar along certain dimensions. In the case of speech perception, this occurs when phonology is added to lexical space.

A potential criticism of these findings is that they have all been based on a distance measure. Although this has been the dominant tool for exploring distributed representations, it is possible that some other measure would be more discriminating. In particular, it may be more useful to examine sparse representations by looking at the angle between the relevant vectors. A reanalysis of the sparseness investigation did remove the comparative disadvantage found for the more sparse representations, but if anything it emphasized the gulf between localist and distributed representations in terms of their capacity to represent activation patterns in parallel.

An alternative is that lexical space should not be treated uniformly, so for example, parallel activation of cohort members may reflect only the phonological dimensions, which are more able to distinguish cohort from non-cohort members and are more interpretable when partially activated. Similarly, given the freedom to construct their own distributed space (e.g., in the hidden units), connectionist networks can ensure that words that are frequently coactivated (such as cohort competitors in speech perception) have similar representations. However, the distributed space must also be able to accommodate unlikely or infrequent combinations of items (perhaps, for example, to entertain the notion of a concrete cow). Also, the distributed space may be subject to separate constraints that do not allow such reorganization of representations. Some of the strongest evidence for early multiple activation in speech perception comes from experiments involving semantic priming (Zwitserlood & Schriefers, 1995). If distributed models are to accommodate these data, then the domain of multiple activation must be a distributed semantic space, which by definition does not permit clustering on the basis of phonological form.

At the moment, the experimental data on the extent of parallel activation in speech perception are equivocal. We do not know how many lexical representations can be activated in parallel, nor whether the number of representations activated affects their degree of activation. Maybe the most profitable reaction to this finding is to accept it as a limitation of distributed connectionism and conduct further experiments to see whether it corresponds to a similar property of the human system. Connectionist models are often accused of being too powerful, but here we have a clear case of something distributed representations find difficult. If this limitation turned out to be one that human systems share, it would be a powerful argument for the validity of modeling cognitive processes using the distributed metaphor.

Acknowledgments

This research was supported by a UK MRC program grant awarded to William Marslen-Wilson and Lorraine Tyler. I am grateful to William Marslen-Wilson, John Bullinaria, Matt Davis, Jeff Elman, Mary Hare and Gary Cottrell for useful discussions of this work.

References

Gaskell, G., & Marslen-Wilson, W. (1995). Modeling the perception of spoken words. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 19-24). Mahwah, NJ: Erlbaum.

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

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, 32, 474-516.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660-665). Mahwah, NJ: Erlbaum.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71-102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.

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

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

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

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32, 25-64.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes*, 10, 121-136.