COGNITIVE SCIENCE Vol 23 (4) 1999, pp. 439–462 ISSN 0364-0213

Ambiguity, Competition, and Blending in Spoken Word Recognition

M. GARETH GASKELL AND WILLIAM D. MARSLEN–WILSON

MRC Cognition and Brain Sciences Unit

A critical property of the perception of spoken words is the transient ambiguity of the speech signal. In localist models of speech perception this ambiguity is captured by allowing the parallel activation of multiple lexical representations. This paper examines how a distributed model of speech perception can accommodate this property. Statistical analyses of vector spaces show that coactivation of multiple distributed representations is inherently noisy, and depends on parameters such as sparseness and dimensionality. Furthermore, the characteristics of coactivation vary considerably, depending on the organization of distributed representations within the mental lexicon. This view of lexical access is supported by analyses of phonological and semantic word representations, which provide an explanation of a recent set of experiments on coactivation in speech perception (Gaskell & Marslen–Wilson, 1999).

I. INTRODUCTION

Several recent models of lexical processing assume that cognitive processing can be treated as dynamic settling activity, where activations of nodes in a network represent relevant perceptual and lexical information (Kawamoto, 1993; Masson, 1995; Plaut & Shallice, 1993). In these models there is no single node representing a single lexical item; instead, word "activations" must be assessed with respect to the pattern of activation across all nodes. In visual word recognition, for example, the input to the network would consist of orthographic features representing the visual input, and the task of the network would be to settle into a stable state across phonological, semantic, and orthographic nodes (Masson, 1995).

Learning new mappings in these models corresponds to altering the network weights to form stable attractor states, and various psycholinguistic measures can be compared to the time the network takes to fall into a particular activation state. For models of this type,

Direct all correspondence to: Gareth Gaskell, Department of Psychology, University of York, Heslington, York YO10 5DD, UK; E-mail: g.gaskell@psych.york.ac.uk

therefore, the final state of the network is only part of the story. The manner in which processing takes place, the intermediate states, and the time (in processing cycles) taken to reach the target state are central part to the account (e.g., Kawamoto, Farrar & Kello, 1994).

Our research on speech perception is closely allied to this approach to cognitive modeling. For speech perception the focus is even more strongly on the course of processing events, rather than just the endpoint. This emphasis is forced by the nature of the stimulus. The speech signal is spread across time, transient, highly variable, and lacks reliable cues to word boundaries. This combination of properties makes states of ambiguity normal rather than exceptional, and necessitates a perceptual system capable of continuous reappraisal of the speech signal, based on partial information.

This sequential processing environment leads to the potential parallel activation of multiple lexical candidates as the speech is heard. Explicitly or implicitly, word recognition involves a process of whittling down of potential candidates (the word-initial cohort) as more sensory input is encountered. Where there is insufficient information to reduce this set to one, there is evidence suggesting that some or all of the meanings of the remaining candidates are activated (Marslen–Wilson, 1987; Zwitserlood, 1989).

The connectionist models we have mentioned treat ambiguity by activating a "blend" of the relevant distributed representations (Smolensky, 1986). Because each word representation involves setting an activation level for every representational node, multiple representations cannot be activated without interference. The resulting activation pattern is a weighted average of the relevant constituents, which may bear some similarity to each of those patterns, depending on various factors. The apparently inescapable interference between distributed word representations contrasts with some localist models, where each lexical item is represented by the activation of an independent node (Morton, 1969; Marslen–Wilson & Welsh, 1978). The contrast between local and distributed models of coactivation becomes crucial in the case of speech perception, where states of transient ambiguity form a dominant part of the perceptual process.

This paper explores the ability of distributed connectionist networks to accommodate parallel activation through blending. We take as a starting-point the basic properties of our network model of speech perception—the distributed cohort model (Gaskell & Marslen–Wilson, 1997-b)—and describe simulations that quantify the model's predictions with respect to parallel activation. To allow precise exploration of many variables, these simulations dispense with connectionist networks and use idealized statistical analyses, based on randomly generated or corpus-derived vector populations. These analyses evaluate the effects of parameters like sparseness on the capacity to activate distributed representations in parallel. They then explore the effects of employing more realistic representations, which encode phonological and semantic similarities between words.

The analyses show that standard distributed network models of cognitive processes are generally poor at supporting the parallel activation of multiple word representations. However, later simulations show that the degree of coactivation depends critically on the organization of the representational space, such that dimensions in phonological space are substantially more suitable for accommodating multiple cohort competitors than semantic

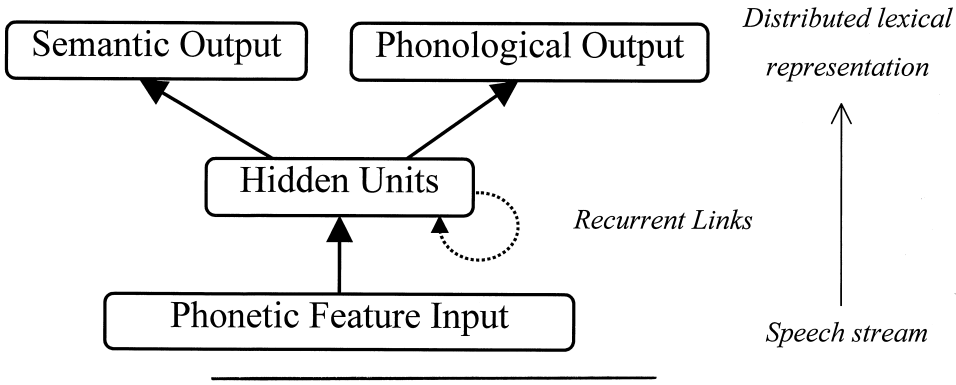

Figure 1. The Gaskell and Marslen–Wilson (1997) distributed connectionist model of speech perception.

dimensions. The final analyses simulate experiments specifically designed to test the predictions of the distributed cohort model. The simulations successfully model the behavioral disparity between good coactivation of phonological representations and poor coactivation of semantic representations during the perception of a word.

II. A DISTRIBUTED MODEL OF SPEECH PERCEPTION

Our research builds on other models of speech perception and word recognition, such as Cohort (Marslen–Wilson & Welsh, 1978; Marslen–Wilson, 1987), TRACE (McClelland & Elman, 1986), and SHORTLIST (Norris, 1994). These are essentially localist, logogen-type models (Morton, 1969), in which the goodness-of-fit between each word candidate and the incoming speech is represented by a separate activation value. Gaskell & Marslen–Wilson (1997-b) examined the effects of implementing lexical access in a distributed system, in which lexical phonological information is an output of the system, along with semantic information. We trained a simple recurrent network (Elman, 1991) to map from a stream of phonetic features onto distributed representations encompassing the meaning and phonological form of words (see Figure 1). Such a model can be thought of as localist at a featural level if consistent meanings are assigned to the individual node values making up the word representation. Alternatively, the lexical representations may be context sensitive "all the way down" (Clark, 1993), with no fixed meanings attached to node values. Either way, the critical issue is that the model contains no equivalent of the word nodes found in localist models such as TRACE or SHORTLIST, and so is distributed at the word level.

Lexical access in this model is interpreted as a trajectory through a high-dimensional space, with the position at any time defined by the activations of the output nodes (the activation of each node corresponds to a position along a separate dimension in the space). Lexical items are represented by relatively stable points and can be thought of as the

desired endpoints for the trajectory. As speech information enters the network, the activation of matching words is reflected by constructing a blend of their representations. When a word onset is presented, the network outputs a blend of the representations of all words matching that onset. As more speech comes in, this blend is refined to represent the reduced set of words that still match the speech. This refinement continues until just one word matches the input. At this point (the uniqueness point, or *UP*) the network can isolate the full distributed representation of the remaining word.

A number of points about this method of modeling speech perception are worth noting. Other models have taken a more serial and hierarchical approach to the representation of different types of information. For example, TRACE (McClelland & Elman, 1986) employs an initial featural level of representation, which feeds into a phoneme level, followed by a localist word level that in a fuller model would presumably map onto one or more levels incorporating lexical information such as word meaning (see Christiansen & Chater, this issue, for further discussion). The distributed cohort model maintains a less hierarchical structure, in which a representation of speech is mapped directly onto lexical representations of form and meaning without undergoing a preliminary stage of categorial labeling as phonemes or similar units.[1]

An advantage of this approach is that it preserves subphonemic detail throughout lexical access rather than integrating information into larger sublexical units (Andruski, Blumstein & Burton, 1994). This property allowed us to simulate data from lexical and phonetic decision experiments (Marslen–Wilson & Warren, 1994), which specifically addressed the nature of the input to lexical analysis, and argued that featural information was not integrated pre-lexically to form phonemic labels.

The location of phonological nodes alongside semantic nodes also allows lexical representations of phonological information to be abstract and dissociated from the acoustics of speech, while retaining the sensitivity of the semantic mapping to fine-grained information. This makes the system intolerant of even minor random deviations in the form of words, while accommodating major phonological changes provided they occur as the result of regular variation in connected speech. This view of lexical phonology is supported by experiments on the perception of phonological changes caused by assimilation of place of articulation in English (Gaskell & Marslen–Wilson, 1996, 1998). We have shown previously, using network simulations, that the perceptual system develops elements of both representational abstraction and contextual sensitivity through exposure to normal variability in the surface form of speech (Gaskell, Hare & Marslen–Wilson, 1995).

A final aspect of the current model is its probabilistic nature. Recent experimental and computational research has demonstrated the value of statistical and distributional information, in both the development of language abilities, and the adult system (Brent & Cartwright, 1996; Cairns, Shillcock, Chater & Levy, 1997; Saffran, Aslin & Newport, 1996). The use of a simple recurrent network allows the model to pick up statistical information during training and to reflect conditional probabilities during states of ambiguity (e.g., before a word's *UP*). The network learns to bias its output during states of ambiguity towards more frequent word candidates (Gaskell & Marslen–Wilson, 1997-b),

reflecting the greater number of training instances involving those words. Indeed, both interactive activation and recurrent network models can be classed as imperfect ways of implementing a purer probabilistic model, that bases word "activations" on conditional probabilities derived from previous experience.

III. BLENDING AND PARALLEL ACTIVATION

The remainder of this article focuses on one prominent aspect of the model—its ability to simulate parallel activation during word recognition. The strongest evidence for parallel activation comes from priming studies using ambiguous speech stimuli (Marslen–Wilson, 1987; Zwitserlood, 1989). Zwitserlood (1989) showed that a spoken word fragment (e.g., /kæpt/ spliced out of the word *captain*) would facilitate timed lexical decisions to visually presented words associated with two potential continuations (e.g., *ship* related to *captain*, *guard* related to *captive*). It seems aspects of both words' meanings are accessed while speech remains ambiguous and that these facilitate related target words. When faced with transient ambiguity, the perceptual system does not wait until speech is uniquely identifiable before accessing lexical semantic information. Other experiments have showed similar priming patterns under a variety of conditions (Marslen–Wilson, 1990; Zwitserlood & Schriefers, 1995).

Localist models accommodate this property easily. These models use the activation metaphor to indicate the status of the recognition process. The degree of match between each word and the incoming speech is reflected in the word's activation value. Localist models can therefore accommodate parallel activation simply by increasing the activation of the relevant word nodes. This means that coactivation can occur without cost in terms of reduced activation, because of the independence of word representations. Regardless of the number of lexical matches to a speech fragment, the nodes representing each lexical item can be independently fully activated (e.g., Marslen–Wilson & Welsh, 1978). However, this property is not obligatory in localist models. TRACE (McClelland & Elman, 1986) and SHORTLIST (Norris, 1994) use inhibitory links between word nodes to implement direct competition between word candidates. Nonetheless, all localist models offer a simple mechanism for representing the coactivation of word representations during lexical access. Indeed, given the importance of parallel activation in speech perception, it is unsurprising that localist activation-based models dominate current theorizing.

In the distributed cohort model, activations are encoded implicitly in the similarity between the network output and each word's distributed representation. In terms of the multidimensional state space through which the network output plots a trajectory, the activation of any word's lexical representation depends on the proximity of that representation to the output of the network. If the output of the network is near to a point representing some word, then that word is said to be activated.

Our simulations investigate this correspondence between distance in the distributed model and localist activations. There are clear differences between the systems; localist models can potentially coactivate multiple word representations simultaneously without cost whereas distributed systems in general cannot (i.e., the output of the network cannot

be identical to two different distributed word representations).[2] We explore how properties of the representational space like sparseness affect the capacity of a distributed system to support coactivation. The aim is to determine whether distributed models provide a plausible alternative to localist models, and to generate a set of predictions from the distributed cohort model, which can then be tested experimentally.

IV. MONTE CARLO SIMULATIONS

Our discussion of the distributed cohort model stated explicitly and precisely how it operates in circumstances of ambiguity, when there is insufficient information in the speech signal to isolate a single matching candidate. The network outputs a blend of the distributed representations of the relevant words. This blend is weighted according to the frequency of the matching candidates. To pursue the *captain/captive* example, if the speech input is/kæptɪ/, and *captain* occurred twice as often as *captive* in the network's training corpus, then twice as many weight adjustments will have biased the network towards the lexical representation of *captain* than *captive*.

Because the network's behavior can be characterized at this abstract level, actually running network simulations to address parallel activation would only add noise. Instead, we use techniques similar to signal detection analyses (Green & Swets, 1966) to investigate parallel activation, based on Monte Carlo simulations with randomly generated vectors. We generate two vector sets with pre-specified properties. These represent a "snapshot" of the recognition process at some point during the perception of a word. One, the "cohort" set, represents the matching word candidates for some speech input, with each vector being a hypothetical distributed representation of a single matching word. The second, "mismatch" set represents all words mismatching the current input. Word recognition in these terms involves setting up a large cohort set based on little or no speech information, and then gradually transferring word representations from the cohort to the mismatch set as more speech is processed. When enough of the speech waveform is available to uniquely identify a word, the cohort set contains one word representation (the identified word) and the mismatch set contains all other lexical items.

Given these two sets, we can then calculate a blend vector—the arithmetic mean of all cohort vectors (weighted according to frequency). This is the idealized output of the network model given the current state of ambiguity. Thus, if the network has learned that its current input could be any of the words in the cohort set, it will produce a blend of all the cohort vectors. We can then examine this blend and ask how well it represents those cohort words. For example, is it more similar to the cohort than to the mismatch set? To address this question, we look at the distributed equivalent of localist activation: distance in lexical space. We can calculate the distance (using an appropriate metric such as root-mean squared [RMS] distance) between the blend and each vector in the cohort and mismatch sets. In localist logogen-type models, an effective model will activate all cohort words above the mismatch words. In the distributed model, an effective blend vector should be nearer the vectors in the cohort set than the vectors in the mismatch set. In signal detection terms, the issue is whether the cohort and mismatch sets are separable along the

dimension of distance from the lexical blend. We can therefore measure the system's effectiveness by examining the degree of separation between different cohort and mismatch populations.

Cohort Size

The first simulation illustrates this procedure by demonstrating the effect of cohort size on coactivation through blending. In this simulation, the "lexical representation" of each word was a randomly chosen 200 component binary vector, with each component having a 50% chance of being set to 1 or 0. The size of the cohort set was varied to simulate different stages of word recognition, when different numbers of lexical items match the speech input. We chose cohort sets consisting of 1, 2, 4, 8, 16, 32, and 64 items to illustrate the behavior of the model. To control for the differing extent of random variation between cohort sets of different sizes, the results were based in each case on 64 values (i.e., for the sets of size 1, 64 analyses were carried out and the results averaged; for the set of size 64, only one analysis was carried out). In each analysis, a blend vector was calculated by taking the mean over all cohort vectors (in this simulation, without frequency weighting). The RMS distance from this blend vector was then calculated for all cohort vectors.

It is important to relate the distance of these lexical blends from cohort representations to the overall population of distances. An effective blend representation of the cohort set should not only be near the cohort vectors, it should also be relatively far from the mismatch vectors. We also examined the distances between each blend and a set of vectors representing mismatching words in the network's mental lexicon. These were 3000 vectors generated randomly using the same procedure as for the cohort vectors. The data are plotted with decreasing cohort set size along the x-axis, illustrating the reduction of the cohort set during the perception of a spoken word (see Figure 2).

When there are many cohort patterns, the signal and noise are merged and the blends are uninformative. At word onset, when the speech signal is highly ambiguous, the model predicts that it will be impossible to decide on the basis of proximity which of the words the blend is intended to represent. However, when the cohort set size is reduced, lexical blends become much closer to their constituent cohort vectors and further from the mismatch vectors. For example, when the blend is based on two cohort patterns, the RMS distance between these patterns and the blend is 0.36—comfortably closer than the nearest mismatch word.

Clearly, modeling parallel activation in this way imposes a limit on the number of words that can be informatively activated in parallel. If too many distributed patterns are blended together, they interfere strongly and there is a good chance of some spurious pattern falling closer to the blend than some cohort patterns. This behavior is a basic statistical property of sampling: the greater the sample size, the better the estimate of population mean. For our purposes, the population mean is the least informative blend state, because it is likely to be equally close to vectors representing cohort and mismatch set members. Hinton & Shallice (1991) show that a blend of two vectors in this type of system is as close to those vectors as any other vector (if not closer). For blends of more

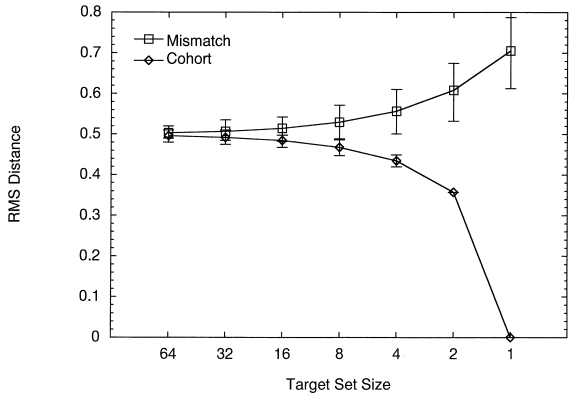

Figure 2. The effect of varying cohort set size on the distance between the cohort blend and the cohort and mismatch populations. For each population, the mean, maximum and minimum distances are marked, with cohort set size plotted on a reversed log scale.

words this does not hold: it becomes possible (and even probable) that other vectors will fall closer to the blend than some cohort vectors.

This result confirms that a distributed system cannot match the ability of localist models to represent the activation of multiple word candidates early in the processing of a speech stimulus. When representations are localist, these candidates can be simultaneously activated without any danger of confusing the active candidates with the inactive ones. The situation for distributed models is more complicated. The simulation does not imply that distributed networks will fail to resolve widespread ambiguity, but it shows that the activations of lexical nodes may not fully dissociate words that currently match the speech input from words that do not. In localist terms, this is like having a noisy model that early on in processing activates some non-cohort members more than some cohort members.

In conclusion, parallel activation of fully distributed representations is possible, but only to a limited degree. This places restrictions on the kinds of experimental results that a distributed model could accommodate. Existing research showing coactivation of lexical representations does not tell us whether parallel activation of multiple lexical representations is complete or only partial (Marslen–Wilson, 1987, 1990; Zwitserlood, 1989; Zwitserlood & Schriefers, 1995). This is an important question because distributed models cannot accommodate complete coactivation of competing lexical representations.

The initial simulation provides a basic picture of coactivation in the distributed model. The following sections explore the extent to which various lexical properties affect the capacity of the system to support coactivation.

Sparseness and Dimensionality

Many models (e.g., Hinton & Shallice, 1991; Plaut & Shallice, 1993) have assumed that distributed lexical representations are *sparse*, meaning that each word's representation

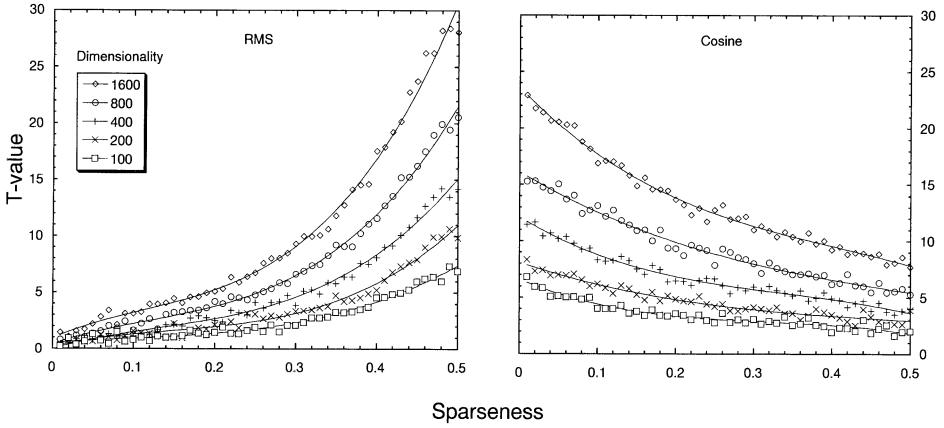

Figure 3. The effect of sparseness and dimensionality on the separability of target and competitor populations using either RMS distance (left hand graph) or Cosine distance (right hand graph). In all cases, the *t* statistic for the target and competitor distances from the blend vector is the dependent variable. The curves are the cubic best fit lines for each level of dimensionality.

will involve the activation of only a few nodes. This factor seems bound to affect the capacity for coactivation—after all, the localist position, which is ideally suited to parallel activation, occupies one end of the sparseness continuum. The representations examined so far, in which 50% of all components were randomly set to 1, lie at the other end of this continuum.

We examined the effect of varying sparseness for cohort and mismatch sets consisting of 50 vectors each. These vectors were generated randomly with a fixed probability of any component being set to 1 within any simulation. However, this probability varied across simulations from 1% to 50% in units of 1%. A second independent variable was the dimensionality of the space (i.e., the number of components in the vectors), which was set to 100, 200, 400, 800 or 1600. For each combination of sparseness and dimensionality, cohort and mismatch sets were generated, and their distances from the blend vector calculated. The separability of the two distributions is summarized using the t-statistic. A high *t* value indicates that the blend vector is generally closer to the cohort vectors than to the mismatch vectors. However, even a very high *t* value does not imply that the two distributions are fully separate, since their tails may still overlap to some extent.[3] Results are plotted for both RMS and cosine distances (see Figure 3). The cosine measure was included because it was expected to be more sensitive than RMS distance for the sparse measures.

The overall effect of increasing dimensionality is clear and unsurprising. The higher dimensional spaces provide a better basis for separating cohort and mismatch sets using either distance measure. In order to discriminate between cohort words (the "active" words) and the mismatch set (all other words in the mental lexicon), the cohort members must be nearer the blend in the lexical space than the mismatch words. If there are few dimensions then the lexical space will be relatively crowded and there is a good chance

of at least one mismatch item being sufficiently close to a lexical blend to cause interference. As the number of dimensions rises, this likelihood diminishes and the capability of the lexical system to accommodate multiple representations increases.

The effect of sparseness and its interaction with dimensionality is more complex. Looking first at the results using RMS distance, there is a strong effect of sparseness on the separability of the cohort and mismatch sets. The distributed representation of multiple lexical items deteriorates as sparseness increases (towards the left of the graph). For very sparse representations, the separability of the sets becomes so low that the two sets are indistinguishable, in that the *t* value is not significantly higher than would be expected by chance. Localist representations are excluded because the *t* value for such a system without noise is infinite. Even allowing for some noise, there is a big difference between the results for a localist system and a sparse system in terms of representational capacity.

The results for the same simulation using the cosine metric are in certain respects quite different. For low dimensionalities the effect of sparseness is small, with sparser representations producing more separable cohort and mismatch sets. However, for high dimensional spaces the advantage of sparse representations is clear, with *t* values close to the peaks reached using the least sparse representations and RMS distance.

The pattern of results using RMS distance can be explained using an extension of the dimensionality argument. Generally, increasing sparseness in a distributed representation deepens the problem of coactivation, despite the fact that the sparser representations seem more similar to a localist representation. Sparse representations are problematic because they place a restriction on the positions in lexical space that words can occupy. This is like reducing dimensionality, which also reduces the capacity of the system. However the very extreme of sparseness—the localist system—is crucially different. It restricts the lexical space but also 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.

To see why the cosine measure works better for sparse representations, we need to look in more detail at how the distances are calculated. In a sparse system, word representations consist of a few ones and many zeros. The blend vector will therefore consist of some zeros and some near zero components (the result of averaging a few ones and lots of zeros). The advantage of cohort vectors over mismatch representations is that the components set to 1 in a cohort vector are guaranteed to be nonzero in the blend vector, whereas the corresponding components in a mismatch vector will be non-zero in some cases and zero in others. This advantage is greater in the cosine measure, because it doesn't matter that the value of the blend component is far less than the value of the cohort component in these cases–only the angle of the two vectors is calculated. This advantage is small when calculating RMS (or any other Minkowski) distance because a small non-zero value is only slightly closer to the target value than zero, and the small advantage is obscured by noise.

But what do these mathematical diversions tell us about the focus of our research: the coactivation of word representations in a distributed system? Increasing the dimensionality of the representational space clearly reduces the problems of coactivation, but it is

difficult to determine where the human system lies along this continuum. 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 to the representation and more obliquely increases the capacity of the system to coactivate multiple lexical entries.

Perhaps then the solution is to ensure that lexical representations are rich. However, there may be quite separate constraints operating that restrict the dimensionality of these representations. Landauer & Dumais (1997) proposed a method for determining the optimal dimensionality of distributed word representations. They examined the extent to which dimensionality reduction of matrices based on word co-occurrence statistics affected the similarity structure of the resulting word representations. They found a peak in performance at roughly 300 dimensions, with fewer or more dimensions merely serving to obscure the similarity structure of many words. If their characterization of word learning is more generally applicable this would impose a relatively low upper limit on coactivation, at least at a semantic level.

Turning to the evaluation of sparseness, we find that sparser representations are better served by the cosine measure, whereas less sparse representations are better served by the RMS measure. It is unclear which of these is most relevant to the mechanics of activation in distributed systems. RMS distance is more obviously comparable to the process of changing node activations, because the actual deviation on each node is considered in the calculation of distance, whereas only the angle between two vectors is relevant to the cosine measure.

Leaving these differences aside, the system that performed best in terms of ease of separation of cohort and mismatch sets was in fact the least sparse system, which gave a t value of over 28 for the largest space using RMS distance. This argues against any attempt to improve the ability of a distributed system to coactivate words by making representations sparser, or near-localist. Any reduction in the overlap between word representations comes at a cost of increased interference between coactive representations.

V. ENCODING STRUCTURE IN LEXICAL SPACE

Semantic Structure

So far, we have assumed a random distribution of word representations in lexical space. However, many researchers assume that distributed mental representations support some kind of similarity structure or "semantic metric" (Clark, 1993) such that items with similar "meanings" have similar representations. The distributed cohort model assumes that lexical representations reflect similarities and differences between words in terms of phonology and word meaning. Each provides structure, which shapes the lexical space and may affect the blending of representations as speech is perceived.

First, we will address the effects of semantic structure on coactivation of distributed representations. Many connectionist models have built semantic structure into representations by selecting a set of simple features that words can be rated on (e.g., *does X have*

legs?), resulting in a binary vector for each word. However, the choice of features, whether hand picked by the experimenter (Plaut & Shallice, 1993) or selected by a panel of subjects (McRae, de Sa & Seidenberg, 1997) is largely arbitrary, and makes it difficult to assess how well they reflect true lexical organization. A rapidly growing body of research has focused on the proposal that semantic similarity can be captured using co-occurrence statistics drawn from large language corpora (Burgess & Lund, 1997; Landauer & Dumais, 1997). This relies on the assumption that words with similar meanings will occur in similar contexts, and has the advantage of automatically generating large sets of distributed representations, which are proving to correlate well with experimental psycholinguistic data on semantic representation.

For these reasons, we analyzed the effects of semantic organization on coactivation using an automatically generated set of co-occurrence vectors, taken from Lund, Burgess & Atchley (1995). There were 2779 word representations with components ranging in value from 0 to 645. Each value signified the number of times the represented word occurred in the context of another, specified word in their corpus. The 200 dimensions that accounted for the most variation were selected, with distances in this space normalized for comparison with the binary spaces. On the assumption that cohort members are semantically unrelated (i.e., excluding morphological relatives and words like *glisten* and *glimmer* that have common semantic features), the cohort word representations were selected randomly from this set, with all other vectors representing mismatching words. The control condition used a random space of 200 binary dimensions, with each component having a 50% chance of being set to 1. As in the initial simulation, we varied the size of the cohort set between 1 and 64, with all other vectors used for the mismatch set. However, in this simulation we defined the separability of the cohort and mismatch sets as the difference between the mean cohort distance and the minimum mismatch distance from the blend. This gave a simple measure of the representational effectiveness of the blend, and unlike *t* values was informative when the cohort set size was small. 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 cohort patterns in parallel. A separability value of zero or less indicates that the two populations are intermixed and that the system is working less well.

We expected the structured representations to differ from the random controls in two ways. First, we assumed they would reflect semantic similarities between words. Second, the dimensions were more continuous, compared to the binary vectors examined so far. To tease apart these two factors, a third analysis used a binary form of the co-occurrence vectors, in which each component was set to either 1 or 0 depending on whether it was above or below the mean value across all words.[4]

Figure 4 displays the effects of coactivation as the size of the cohort set is reduced. Both forms of structured vectors suffer more from coactivation than the random vectors. The zero crossing for the random vectors is at roughly 16 patterns, whereas for the binary structured vectors it is under 4, and for the continuous structured system it is below 2. Thus, 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.

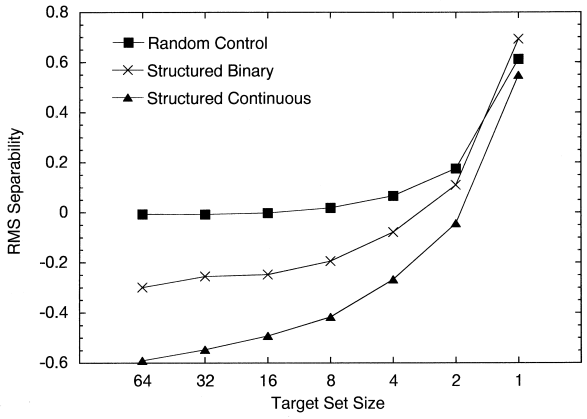

Figure 4. Effect of semantic clustering on separability for three models of semantic representation. The separability scores for the continuous space are normalized to facilitate comparison with the other distances.

The more realistic space creates problems distinguishing signal from noise because groups of words form tight clusters in the space. Representations of food words, for example, may be similar to each other but very different to all other representations. When one of these is blended with the representation of an unrelated word there is a good chance that another word in the cluster will be as close or closer to the blend than the cohort member. In terms of using proximity to separate cohort and mismatch populations, semantic clustering causes problems. However, in some cases the important question is not whether cohort and mismatch populations are discriminable. Instead, one might ask whether there is any useful semantic information encoded in the blend. This is the case in semantic priming, where the target activation is compared to a single (usually representative) member of the mismatch set. We return to this issue in the final simulation.

The non-binary form of the structured representation fares even worse because there are no restrictions on the positions word representations can occupy in the space. In particular, words may occupy positions near the middle of the space, which is where the blends, being arithmetic means, tend to sit. The partitioning of the binary spaces into blend states near the center and single word representations at the "corners" of the space (cf. Anderson & Mozer, 1981; Plaut, 1997) can be useful, because it provides a simple, explicit metric for determining the state of the recognition process without having an external decision mechanism requiring information about which states in the space correspond to words.

Phonological Structure

The previous simulation indicated that adding realistic clustering generally worsens the problem of activating distributed representations simultaneously. However, there is one case where realistic clustering lessens this problem. For speech perception this is the case

where lexical dimensions reflect similarities in the phonological form of words, as in the distributed cohort model (see Figure 1). This is because the phonological representations of words that must be activated in parallel (i.e., cohort members) are guaranteed to be more similar to each other than to unrelated words. Along the dimensions that encode the similarities, the blend will match the cohort representations exactly, but will mismatch other words. This gives the cohort set a head start in terms of overall distance to the blend in lexical space, and decreases the chances of non-cohort members falling close to the blend vector.

This coherence leads to strong predictions about the effects of coactivation in different subsections of lexical space. We assume that during the course of hearing a spoken word a blend of the matching lexical representations is built up and continuously modified based on the uptake of new information. This blending process will take place across both semantic and phonological nodes, with coactivation on phonological nodes generally more coherent than on semantic nodes (because of the phonological similarities between the coactivated words).

These effects of coactivation in different areas of lexical space can be teased apart by examining the facilitation of different types of target words in cross-modal priming experiments. We assume that priming in a distributed model of lexical processing depends on the similarity between the relevant words' representations (cf. Burgess & Lund, 1997; Masson, 1995). A prime word will facilitate recognition of a target word to the extent that its lexical representation is more similar to the target representation than an unrelated baseline—in effect, the prime representation enjoys a proximity advantage over the control representation. In repetition priming, the target lexical representation is related to the prime representation in all dimensions, so recognition of the target can take advantage of overlap on both semantic and phonological nodes (for a prime stimulus consisting of a word fragment, the coherence on phonological nodes will be particularly useful in the recognition of the target word). In contrast, semantic priming relies on overlap on the semantic nodes alone, so any coherence built up on the phonological nodes is unable to facilitate responses.

Experimental Data. We will demonstrate the effects of coherence on coactivation in phonological and semantic lexical space by simulating a set of experiments designed specifically to test the predictions of our model, and developed in parallel with the modeling work. The experiments (see Gaskell & Marslen–Wilson, 1997-a, 1999, for a fuller description) were a development of previous research on parallel activation, using cross-modal priming to measure the activation level of lexical items as the speech signal is heard (Zwitserlood & Schriefers, 1995). We used bisyllabic auditory prime words, which were presented either complete (e.g., *captain,* pronounced/kæptɪn/), or in two splice conditions. At Splice 1 (see Table 1) the final vowel and consonant(s) were removed (e.g., /kæpt), and at Splice 2 just final consonant(s) was/were spliced out (e.g., /kæptɪ/). These fragmented primes were intended to generate different levels of competition or ambiguity in the perceptual system, reflecting the different numbers and frequencies of words that transiently match the speech input during the course of spoken word recognition. Com-

TABLE 1
Priming Experiments: Design, Stimuli, and Results

	Early UP			Late UP		
	Splice 1	Splice 2	Complete	Splice 1	Splice 2	Complete
Test prime	''garm''	''garme''	''garment''	''capt''	''captai''	''captain''
Control prime	''chis''	''chise''	''chisel''	''mount''	''mountai''	''mountain''
Conditional probability	0.75	0.95	1.00	0.24	0.44	1.00
Repetition priming (ms)	35**	78**	92**	36**	47**	64**
Semantic priming (ms)	12(*)	28*	22**	3	10(*)	15*

Note. Conditional probability measures are given for each prime condition (see text), followed by the mean priming effects (control-test) for the repetition priming and semantic priming experiments. The priming values are marked according to their statistical significance, based on the least significant *F*-ratio of the items and participants analyses: **, $p < .01$; *, $p < .05$; (*), $p < .1$.

petitor environment was also manipulated by choosing prime words with varying sizes of cohort sets and varying uniqueness points. These are referred to in Table 1 as the Early *UP* and the Late *UP* conditions. In the Early *UP* conditions, the prime words became uniquely discriminable from other cohort members by Splice 2. In the Late *UP* conditions, the primes became unique at the ends of the words.

The manipulations of prime fragment length and competitor environment were intended to create sets of word fragments that varied widely in the number and frequency of the matching lexical items. To assess more formally the degree of ambiguity, we calculated a conditional probability value for each prime stimulus. This was the CELEX database frequency (Baayen, Piepenbrock, & van Rijn, 1993) of the complete prime word, divided by the summed frequencies of all morphologically unrelated words matching the stimulus in terms of phonemic representation. The result is an estimate of the probability of the given prime stimulus turning out to be the complete prime rather than some other member of the same cohort. A value of 1 on this scale implies that the prime stimulus unambiguously matches a single lexical item, whereas a value close to zero implies either many cohort competitors or a few high frequency competitors. Table 1 includes the mean probability estimates for each prime condition.

To measure the effects of these variations in competitor environment at different splice points, a related visual target was presented immediately at prime offset. Activation of the lexical representation of the complete prime word was assessed by measuring the speed of response to the target, compared to the response time to the same target when preceded by an unrelated control prime. The standard assumption here is that the degree to which recognition of the target is speeded by the related prime (compared to the control condition) reflects the degree to which the prime's lexical representation has been activated. Crucially, the experiments used two types of prime-target relationship, as described below.

According to our model, the response of the perceptual system to these prime stimuli is to generate a frequency-weighted lexical blend, made up of the distributed representations of the phonological form and meanings of all matching words. We predicted that because of the greater coherence between the cohort members' phonological representa-

tions, coactivation would operate more effectively on the phonological nodes than the semantic nodes. We tested this across experiments by using two types of visual target, expected to be sensitive to overlap of different types of lexical information. In the repetition priming experiment the target was the orthographic form of the complete prime word (e.g., CAPTAIN). We assumed this target would access the same modality-independent lexical representation as the complete prime, so that the pre-activation of the full target lexical entry (semantic and phonological) by the prime would facilitate responses to the target. Where the prime was fragmented, the facilitation would depend on how similar the resultant lexical blend was to the target representation both phonologically and semantically. The fact that phonological information was shared between prime and target in this case meant that the effects of increasing competition (or ambiguity) on priming should be weak, because of the coherence of the coactive words' phonological representations.

In the semantic priming experiments the target was related to the prime in meaning, but not in form (e.g., *captain*-COMMANDER).[5] In this case the state of the lexical blend in terms of the phonological representation generated should not affect target recognition, since it is unrelated to the prime word in terms of form. Instead, the only basis for priming is semantic, which we have argued does not provide a sound basis for parallel activation of many cohort members. We expected that the effects of competition on priming in this experiment would be severe, because highly ambiguous prime tokens would create an uninformative blend at the semantic level, which would fail to facilitate target recognition.

To summarize, the experiments aimed to examine coactivation and blending of lexical representations across the whole distributed vector (using repetition priming) and across the semantic vector alone (using semantic priming). Facilitation in the semantic priming experiments (the response time advantage when the target is preceded by the related prime as opposed to the unrelated control) should be weaker and more prone to competition effects than in the repetition priming experiment.

The results of the three experiments were analyzed in two ways. First, analyses of variance were carried out, based on the factors of competitor environment (Early/Late Uniqueness) and prime fragment length (Splice 1, Splice 2, Complete)—see Table 1 for condition means. For the repetition priming experiment, the facilitatory effect of the related primes was robust for all combinations of *UP* and prime fragment length, but the amount of priming in each case was affected by both prime fragment length and competitor environment. Longer primes with fewer competitors produced the largest priming effects. Facilitatory effects in the semantic priming analyses were far smaller, and were marginal or non-significant for the conditions containing the most ambiguous primes (the late separation stimuli at the first two splice points and the early separation stimuli at the first splice point). Competition effects were evident in the significant effect of competitor environment on priming.

The second form of analysis examined the overall linear correlations in each experiment between the degree of ambiguity associated with each individual test prime and the amount of facilitation found (see Figure 5). We will focus on these analyses as they are more closely comparable to the simulations reported here.

Comparing the linear plots for each experiment, the most obvious result is that

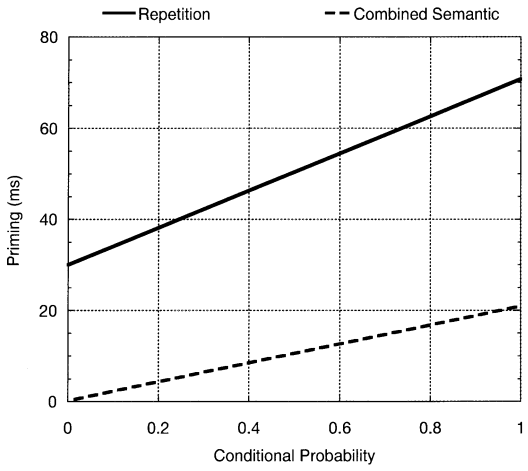

Figure 5. Summary of the Gaskell and Marslen–Wilson (1999) results. The lines plots the correlation between the conditional probability attached to each prime stimulus and the amount of facilitation (unrelated RT - related RT) for repetition (solid line) and semantic (dashed line) priming. The plot of the semantic priming data combines the results from two experiments.

repetition priming effects were much larger than semantic priming effects. This is predicted by almost any model of lexical access, and is relatively unimportant. The critical issue is the effect of ambiguity or competition on the basic priming effect. Does increasing the ambiguity of the stimulus gradually eliminate priming, or does some residual priming remain? To examine this question we looked at the y-axis crossovers. Here we found a dissociation between the two types of priming. In the semantic experiments, facilitation dropped to zero as the conditional probability dropped to zero. Using the combined semantic priming data, there was a significant correlation between conditional probability and facilitation and no significant constant factor in the regression equation for these variables. It seems that semantic priming directly reflects conditional probabilities. The degree to which any prime facilitated recognition of a semantically related word was proportional to the probability of the stimulus actually being that prime word, based only on the speech information contained in the fragment.

The competition effect for the repetition stimuli was different. As before there was a significant correlation between conditional probability and facilitation, but there was also a significant constant in the regression equation. This is evidence of weaker competition, in that the facilitation did not drop to zero with conditional probability. In states of strong competition or high ambiguity (towards the left side of the graph) there was nonetheless some facilitation of the target word. We argue that this support comes from the coherence of the cohort members' phonological representations, which provides a basis for their coactivation, facilitating the activation of the target word.

Simulation. The pattern of experimental results broadly agrees with the predictions of the distributed cohort model. Here we model these findings in detail, using blending rather than connectionist simulations to provide a clearer analysis of the fit between theory and

data. The procedure for this simulation is slightly different to the preceding ones. Previously, we have used large sets of "mismatch" vectors as our baseline for examining the effectiveness of coactivation. Here we simulate the experimental situation, where a related prime is compared to a control prime in the extent to which it facilitates recognition of a single target word. For this reason we compared a single target to two blends, representing the lexical activation caused by hearing either the related prime or the unrelated control. If the related prime blend is closer to the target position than the unrelated control, then priming should occur (cf. Plaut, 1995; Burgess & Lund, 1997). To perform this simulation accurately we need plausible semantic and phonological representations of all words relevant to the experiment (i.e., 42 test primes, 42 control primes, and 42 semantic targets, as well as 697 cohort competitors of the primes, selected from CELEX to match the phonemic representation of at least one fragmented test prime).

Word meaning was again represented by vectors taken from corpus analyses of co-occurrence data, this time extracted from the British National Corpus of speech (Levy, Bullinaria & Patel, 1997).[6] The vectors were generated from two large co-occurrence matrices, created by calculating the proportion of times the words of interest occurred in the context of the 4100 most frequent words in the corpus. For one matrix, co-occurrences were counted using a window of two words to the left of the target word. For the second matrix co-occurrences were counted in a window of two words to the right of the target word. The 55 columns that accounted for the greatest variation between the rows representing the experimental stimuli were chosen, resulting in a 55 component vector for each word. The values of the components were all between 0 and 1 and were generally close to 0.

Capturing the phonology of the stimuli required a representation that encoded similarities between words of widely varying length. The representation we chose consisted of simply a component for each phoneme in the English language, with the value of each component for any particular word being the number of occurrences of the associated phoneme in that word. An additional component encoded the number of syllable boundaries within the word, bringing the total number of components to 55, matching the dimensionality of the semantic representation. The representation of the word *connectionist* (/kənekʃənɪst/) using this system would have values of 3 on the syllable boundary node (4 syllables, so 3 boundaries), 2 on the /k/, /n/ and /ə/ nodes, 1 on the /e/, /ʃ/, /ɪ/, /s/, and /t/ nodes, and 0 elsewhere. This simple representation is not optimal, because it does not capture order information and so cannot distinguish between pairs like *cat* and *act*. Nonetheless, as a rough measure of similarity between words of varying length it works well.

To check the suitability of the semantic vectors, we calculated the mean RMS distances from the target representation to the related and unrelated prime representations. Within each triplet, the target representation was generally closer to the related than the unrelated prime representation (this was the case for 75 out of 84 triplets). The differences were small, but nonetheless significant (mean distance from related prime = .010, mean distance from unrelated prime = .019; difference = .009; $t = 9.5$, $p < .001$). Thus the co-occurrence vectors correctly predict that semantic priming should occur, based on a lexical proximity model of priming. The experimental data by comparison are noisier: of the 84 mean response time difference scores only 57 had positive facilitation values.[7]

We then conducted two blending simulations, one using only the semantic vectors (semantic priming), and one using the combined semantic and phonological vectors (repetition priming). Our assumption was that semantic priming would depend only on the activation of the semantic nodes because both test and control primes are phonologically unrelated to the target. On the other hand, repetition priming can make use of overlap of the full lexical representations on both semantic and phonological nodes. In each simulation, for each related prime word, three vectors were constructed. One was simply the relevant part of the vector for that word (the full vector for repetition priming or the semantic part only for semantic priming). This vector represented the idealized output of the network after presentation of the complete prime word. The other two vectors were blends, representing the state of the network at earlier points, when less of the speech had been presented and the identity of the prime word was ambiguous. These were frequency-weighted blends of all the word representations that matched the particular fragments used in the experiment. For example, if the prime stimulus was /wɪkɪ/–a fragment of *wicket* that also matches *wicked*–the blend would be the mean of the two vectors for *wicked* and *wicket*, weighted according to relative frequency. This vector corresponds to the expected output of the model at the point where the second /ɪ/ of /wɪkɪ/ is presented as input.

The vectors for the control words were calculated using exactly the same procedure, but with vectors that were unrelated to the target word. We then calculated the distances between all prime vectors and the target vector. For the repetition priming simulation, the target vector was the full vector representing the prime word (e.g., the vector for *captain*), whereas for the semantic priming simulation, it was the vector representing the meaning only of the semantically related target word (e.g., the vector for *commander*). The difference between the test and control prime distances yielded a proximity advantage for each related test item. These results are summarized in Figure 6 in terms of the correlation between this distance and the frequency-weighted competition measure for each item.

Looking first at the results for the full semantic and phonological space (see Figure 6, left hand graph), we found both a significant correlation, ($r = .87$, $p < .001$) and a significant constant in the regression analysis ($t = 19.3$, $p < .001$). The proximity advantage is linearly related to the conditional probability, but *does not* diminish to zero as ambiguity increases. Instead, there is a residual advantage, due to the coherence between the phonological representations of the target word and its higher frequency competitors. This pattern of results is similar to the repetition priming results from our experiments.

When the same simulation was carried out using semantic vectors alone, and with semantically related target words (see Figure 6, right hand graph), we again found a positive correlation between the conditional probability of the prime stimulus and the simulated priming effect ($r = .34$; $p < .001$). Related prime fragments that were less ambiguous showed a greater proximity advantage to the target. The simulation also shows that the proximity advantage diminishes to zero as the conditional probability drops to zero (in a regression analysis, the constant factor was not significant; $t = 0.82$, $p = .41$). This again replicates the experimental findings.

Because the correlation in the semantic priming was weaker, there is a possibility that the greater level of noise in this simulation was obscuring the presence of a constant factor

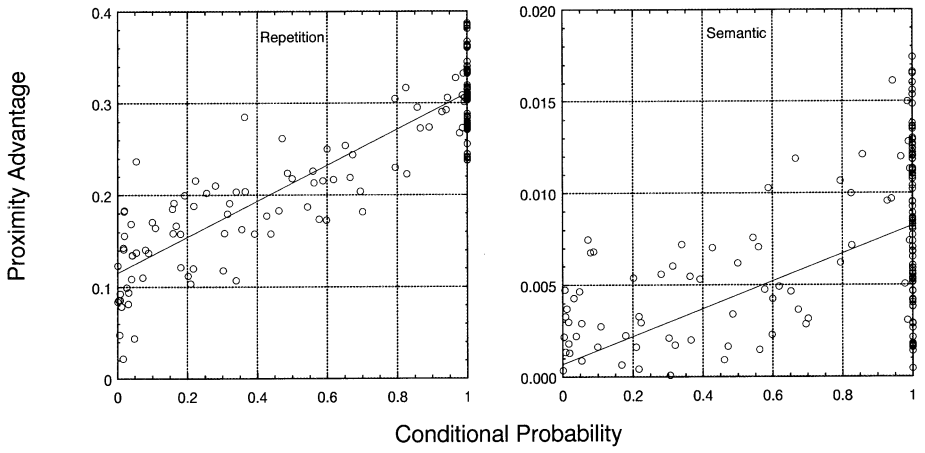

Figure 6. Correlation between the conditional probability associated with each stimulus and proximity advantage for the repetition (left hand graph) and semantic (right hand graph) priming simulations. The circles mark individual data points and the lines illustrate the linear best fit. In order to focus on the best fit lines, the y-axis limits are chosen to exclude a small proportion of the data points (mostly from the bottom of the semantic priming graph).

in the regression analysis. For this reason, 9 replications of this simulation were carried out, with the same set of semantic vectors, but with a random assignation of vectors to control primes (i.e., the representations of the test primes were unchanged, but the control representations were randomly chosen each time. The analysis of the average results across the 10 simulations showed a strengthened correlation ($r = 0.55$, $p < .001$) but still not a hint of a residual constant in the regression analysis ($t = 1.1$, $p = 0.27$).

We have assumed that repetition priming is based on full overlap between prime and target along all dimensions. However, a further simulation showed that much the same pattern of results emerges if repetition priming only makes use of overlap along lexical phonological dimensions. Again there was a significant correlation and a significant constant in the regression analysis.

One further point is worth noting about these data. A potential discrepancy between experimental results and the simulation involves the comparison of absolute distance between experiments. The proximity differences in the repetition simulation are a full order of magnitude greater than the differences in the semantic simulation. This is largely because the semantic vectors contain very small numbers. The actual values of the components are probabilities of co-occurrence, and so are often 0 and seldom higher than .01, meaning that the word representations were all near the origin in the representational space. In contrast, the components of the phonemic vectors had a greater proportion of 1s, with occasionally 2s or 3s. We could easily scale up the semantic vectors or change the dimensionality to reduce the difference between these spaces, but this would simply be a *post hoc* data fitting exercise.

A more valid comparison looks at percentage priming: what proportion of the control-target distance does the proximity advantage comprise? If we examine the best-fit lines in

this way, we find at least a rough correspondence. The percentage proximity advantage for the semantic simulation rises from 5% to 45% as the conditional probability from 0 to 1. Along the same continuum, the repetition simulation rises from 40% to 100%. Therefore the maximum semantic priming is about 1.1 times the minimum repetition priming (0.7 in the experiments) and about half the maximum repetition priming (0.3 in the experiments). However, the fact that the prime-target relationship was varied between experiments with different subjects makes it difficult to get a good comparison of priming magnitudes.

To summarize, we have demonstrated the importance of phonological coherence in the distributed model. This provides a sound basis for coactivation of cohort members, allowing partial activation of many different words on the phonological nodes. The differences between coactivation for distributed semantically and phonologically organized representations allow us to simulate the priming results of Gaskell & Marslen–Wilson (1999) in some detail and provide an insight into the reduced competition effect found using repetition priming.

The qualitative correspondence between the model and the experimental data supports the proposal that overlap along distributed lexical dimensions provides a basis for modeling priming (cf. Burgess & Lund, 1997; Lund, Burgess, & Atchley, 1995). In both repetition and semantic priming, overlap between prime and target along the relevant dimensions can explain the pattern of priming. This is unlike the standard model of priming, particularly in the case of repetition priming, where the prime is assumed to pre-activate a localist representation of the target word. However, we should stress that these results do not rule out the localist representation of word nodes. One possibility is that a localist word level exists below the semantic level, with semantic priming dependent on activations at a semantic level and repetition priming dependent on activations at the word level. Such a model would still need to explain the weakness of competition effects in repetition priming, but a suitable combination of activation, inhibition and resting level parameters could possibly be found. An alternative is that the localist word level resides below the distributed semantic and phonological level that we are proposing, but does not take part in semantic or repetition priming. This would shift the focus of competition effects from a localist word level to the domain of lexical content.

VI. DISCUSSION

This article has addressed a critical test case, not only for our model of speech perception, but for distributed models of language processing in general. Ambiguity is present in many areas of cognition, but speech perception involves a particularly protracted state of ambiguity, when many different words are transiently compatible with the sensory evidence. We have looked at how distributed systems cope with this widespread ambiguity, and at the consequences in terms of activation of lexical information.

There is quite a strict limit on the number of distributed patterns that can be usefully represented by a single blend. In general, more than a handful of representations results in a noisy blend, for which simple distance in lexical space does not properly distinguish the components of the blend (the word-initial cohort) 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.

Various structural factors affect the capacity for representation of multiple lexical items. It correlates positively with the dimensionality or degrees of freedom in the lexical space, implying that the capacity for coactivation of lexical representations will be somewhat greater if a high dimensional space is used. However, this positive effect of increasing dimensionality may have to be balanced against the potential for high dimensional spaces to become noisy in their representation of similarity (Landauer & Dumais, 1997). The sparseness of lexical representations has a more complex effect, but even quite sparse representations do not increase the capacity to accommodate multiple distributed representations, despite their surface similarity to localist representations.

To us, the most interesting result of our simulations is the finding that the organization of lexical representations in the multidimensional space has powerful effects on the capacity of the system for coactivation. With no external constraints on the organization of the space, linearly independent distributed patterns can be chosen, which behave as a localist system (Smolensky, 1986). However, if the structure of the space is required to reflect similarities and differences between words in terms of what we know about them, then this luxury cannot be afforded. Instead, the degree of coherence between the distributed representations becomes all-important. Competition in this type of model is interpreted in terms of activation of lexical content, rather than interaction between abstract word identifiers.

We have characterized speech perception as a mapping onto distributed semantic and phonological representations, although other forms of knowledge may also be activated. What is important in terms of coactivation is the degree of regularity involved in the mapping from the speech wave onto lexical knowledge. The mapping onto meaning is usually arbitrary, so words that are required to be activated in parallel will have incompatible semantic representations, causing strong competition between the distributed semantic representations. On the other hand, lexical phonology has a far more regular relationship with the surface form of speech, providing coherence between the phonological representations of cohort members, and supporting a weaker form of competition on the phonological nodes. These novel predictions receive detailed support from cross-modal priming experiments (Gaskell & Marslen–Wilson, 1997-a, 1999), showing the predicted dissociation between effects of competitor environment on repetition and semantic priming.

Acknowledgments: We thank Curt Burgess, Kevin Lund, Joe Levy, and John Bullinaria for providing corpus-based vectors, and members of the Center for Speech and Language for valuable discussion and comments. Thanks are due also to Morten Christiansen, Nick Chater, and two anonymous reviewers for their comments. This article is based in part on a paper presented at the 18th Annual Cognitive Science Society Conference, San Diego, CA.

NOTES

1. The use of phonological features as input in the current implementation of the model reflects the demands of computational tractability, rather than a theoretical claim about the format in which speech input is input to the lexical system.

2. Note that distributed representations can behave exactly like localist ones in combination, as long as they are unbounded and linearly independent (Smolensky, 1986). Thus, a network may be able to develop distributed representations that are ideally suited for the purposes of parallel activation. However, the case we consider here is one in which distributed representations are constrained by other factors (such as encoding semantic, and phonological similarity) and will generally not be linearly independent.
3. In a small number of cases (for some of the 800 and 1600 dimensional spaces with high sparseness values and using the RMS measure) the cohort and mismatch set distances did not overlap. However, this state of affairs would be unlikely if the mismatch set was of a more realistic size (e.g., 50,000 rather than 50 words).
4. This method of binarizing the vectors is not particularly sensitive to the similarity structure of the representations. Less clumsy techniques exist (Clouse & Cottrell, 1996), but these also increase the dimensionality of the space, introducing a confounding factor.
5. Two separate semantic priming experiments were conducted, varying the delay between the offset of the prime and the onset of the target. Here we report the combined results of these two experiments.
6. The corpus analyses were carried out by John Bullinaria and Joe Levy. We thank them for making their results available to us.
7. Levy et al. (1997) used our semantic priming experiments as one of their yardsticks for a systematic analysis of various parameters involved in corpus analyses, such as window size and type. They found that a range of different windows provided similar degrees of priming, but that certain distance measures were more effective than others (Hellinger distance was the most effective). However, to maintain compatibility with earlier simulations, we continue to use RMS distance here.

REFERENCES

Anderson, J. A., Mozer, M. C. (1981). Categorization and selective neurons. In G. Hinton & J. Anderson (Eds.), *Parallel models of associative memory* (pp. 213–236). Hillsdale, NJ: Erlbaum.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52,* 163–187.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93–125.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 177–210.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation. *Cognitive Psychology, 33,* 111–153.

Christiansen, M. H. & Chater, N. (this issue). Connectionist natural language processing: The state of the art. *Cognitive Science*

Clark, A. (1993). *Associative engines: Connectionism, context and representational change*. Cambridge, MA: MIT Press.

Clouse, D. S., & Cottrell, G. W. (1996). Discrete multi-dimensional scaling. In G. W. Cottrell (Eds.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 284–289). Mahwah, NJ: Erlbaum.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7,* 195–225.

Gaskell, M. G. (1996). Parallel activation of distributed concepts: who put the P in the PDP? In G. W. Cottrell (Eds.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 284–289). Mahwah, NJ: Erlbaum.

Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science, 19,* 407–439.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 144–158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997-a). Discriminating local and distributed models of competition in spoken word recognition. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 247–252). Mahwah, NJ: Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997-b). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Gaskell, M.G., & Marslen–Wilson, W.D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology-Human Perception and Performance, 24,* 380–396.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Representation and competition in the perception of spoken words. Manuscript submitted for publication.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Levy, J. P., Bullinaria, J. A., & Patel, M. (1997). *The evaluation of the use of co-occurrence statistics.* Paper presented at the First International Conference on Computational Psycholinguistics. University of California at Berkeley, CA, July.

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 Seventeenth Annual Conference of the Cognitive Science Society* (pp. 660–665). Mahwah, NJ: Erlbaum.

Marslen–Wilson, W., & Warren, P. (1994). Levels of representation and process in lexical access. *Psychological Review, 101*(4), 653–675.

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

Marslen–Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). Cambridge, MA: MIT Press.

Marslen–Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

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

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

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

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

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

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37–42). Mahwah, NJ: Erlbaum.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science, 274*(5294), 1926–1928.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 390–431). Cambridge, MA: MIT Press/Bradford Books.

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.