LANGUAGE AND COGNITIVE PROCESSES, 1997, *12* (2/3), 177–210

Modelling Parsing Constraints with High-dimensional Context Space

Curt Burgess and Kevin Lund

University of California Riverside, Riverside, CA, USA

Deriving representations of meaning has been a long-standing problem in cognitive psychology and psycholinguistics. The lack of a model for representing semantic and grammatical knowledge has been a handicap in attempting to model the effects of semantic constraints in human syntactic processing. A computational model of high-dimensional context space, the Hyperspace Analogue to Language (HAL), is presented with a series of simulations modelling a variety of human empirical results. HAL learns its representations from the unsupervised processing of 300 million words of conversational text. We propose that HAL's high-dimensional context space can be used to (1) provide a basic categorisation of semantic and grammatical concepts, (2) model certain aspects of morphological ambiguity in verbs, and (3) provide an account of semantic context effects in syntactic processing. We propose that the distributed and contextually derived representations that HAL acquires provide a basis for the subconceptual knowledge that can be used in accounting for a diverse set of cognitive phenomena.

INTRODUCTION

All accounts of syntactic processing, language comprehension and memory posit a role for semantic representations. An important aspect of any model of language comprehension is to specify the nature of the various components of the comprehension system and to what extent these components interact. Psycholinguists have long debated the extent to which semantic information can influence syntactic processing, and at what point in the comprehension process this occurs. Two general views have

Requests for reprints should be addressed to Curt Burgess, Psychology Department, 1419 Life Sciences Building, University of California Riverside, Riverside, CA 92521-0426, USA. E-mail: curt@cassandra.ucr.edu

This research was supported by a NSF Presidential Faculty Fellow award SBR-9453406 to C.B. We thank Ken McRae, Nick Chater, Debra Long, Kay Livesay, Catherine Decker and two anonymous reviewers for their helpful comments and criticisms, and Maureen Keeney for bibliographic assistance.

predominated in this controversy: one is syntactically motivated, whereas the other is lexically motivated. Empirical studies in this domain have relied heavily on the resolution of syntactic ambiguity to marshall evidence for either view.

The syntactically motivated theories presume that syntactic ambiguity is resolved in two stages. First, the syntactic processor (or parser) uses a set of parsing principles to build one or more syntactic constructions. The second stage of processing uses lexical/semantic or discourse-level information that can be used to resolve ambiguity and continue to build a mental model of the ongoing discourse. Although there are important differences in some of these models (Frazier, 1978; Frazier & Fodor, 1978; Gorrell, 1989; Kurtzman, 1985; Rayner, Carlson, & Frazier, 1983), they are in distinct contrast to the more lexically motivated models. These more lexically motivated parsing models, particularly those inspired by lexical ambiguity research (Burgess & Hollbach, 1988; Burgess & Lund, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Carlson, 1989), allow more flexibility for communication between language subsystems and tend to incorporate more information into the lexical/semantic representations that can facilitiate making the necessary commitments that are required during on-line language comprehension. Semantics in these more constraint-based models plays a crucial role in the parsing operation, in that it can provide the parser with guidance from real-world knowledge (for reviews, see MacDonald et al., 1994; Tanenhaus & Carlson, 1989).

Though lexically motivated and syntactically motivated theories differ with respect to issues of processing modularity, they agree for the most part that there is representational modularity. For example, there is a level of semantic or discourse representation that can (or cannot) affect the behaviour of a syntactic processor. In this paper, we present a representational model that seems to mimic effects that are considered (1) semantic, (2) grammatical, or (3) show an interaction of semantic and syntactic information. We argue that the representations for words that are generated with this model are subconceptual in nature and blur the typical representational distinctions that are made.

The aims of this paper are to describe our representational model and to present a set of simulation results that bear on some of these currently debated questions, which concern semantic contraints on parsing and grammatical categorisation. In earlier papers, we have presented the Hyperspace Analogue to Language model, or HAL, as a model of semantic memory. There was good reason to make this argument. HAL provided an account of single-word priming with semantically related items whether they were associatively related (e.g. *cat–dog*) or not (e.g. *beer–ale*). Words that were not category instances, but were associatively related, using word association norms (e.g. *crater–moon*), did not show a relatedness effect using

the HAL model (Lund, Burgess, & Atchley, 1995; Lund, Burgess, & Audet, 1996). Likewise, the model provided an account of semantic paralexias generated by patients with deep dyslexia, whereas word association norms did not (Buchanan, Burgess, & Lund, 1996). However, our more recent research, presented in this paper, suggests that the characterisation of HAL's representations as semantic is too specific and limiting.

HAL acquires its representations from monitoring the contexts in which words occur. HAL uses a set of these lexical co-occurrences as coordinates in a high-dimensional space that can provide a measure of similarity between lexical items. HAL's similarity measure is essentially a measure of contextuality, a notion upon which we will expand later. In brief, we will argue that this measure of contextuality can provide the various types of information that one would want in a rich memory system. The first part of the paper will describe various ways in which semantics have been modelled, followed by a description of the methodology HAL uses to generate its semantic representations. The first simulation provides a demonstration that HAL's representations can be used to make categorical distinctions among semantic categories. At this point, we will discuss the distinction between the vector elements in HAL's representations and the traditional notion of semantic features. The second simulation demonstrates that the ability of HAL's vectors to categorise information extends to basic grammatical categories as well. In the third simulation, we show that the vector representations carry sufficient information to make more fine-tuned grammatical distinctions. The final simulation demonstrates that the interword relationships, as represented by the vectors, correspond to the semantic constraints on parsing that have been investigated in a series of experiments with human subjects.

Representing Semantics

Early work in word recognition and lexical ambiguity builty heavily on localist models of semantic memory. One of the most frequently cited semantic network models is the spreading activation model introduced by Collins and Quillian (1969; see also Collins & Loftus, 1975). In this model, semantic meaning was represented by nodes which corresponded to individual words. Nodes in a spreading activation model are connected by links which represent the type of relationship shared by the nodes; the links also vary in length, reflecting the strength of the relationship. One elementary result of this model is that concepts facilitate the subsequent processing of related concepts.

Meyer and Schvaneveldt (1971) provided an early demonstration of how recognising semantically related words (*bread–butter*) can speed the lexical decision latencies as compared to seeing unrelated words (*floor–butter*).

Semantic memory research using the lexical decision (or naming) task in the last two decades has produced one of the largest bodies of cognitive psychological literature (see Neely, 1991). The presence of the semantic priming effect is one of the most robust effects in the literature, although the exact nature of the word relationships and the nature of the task and methodology can influence the magnitude and presence of priming (Fischler, 1977; Moss, Ostrin, Tyler, & Marslen-Wilson, 1995; Neely, 1991). For some of the parsing experiments that will be discussed later, it is important to note that reading one word has important consequences for the processing of a subsequent word (i.e. priming related concepts and otherwise offering various constraints).

The precise characteristics of the mental representations that underlie the semantic priming effect and the relationship of concepts to one another is an issue with a long history in cognitive science. Early approaches used psychometric techniques to compute a distance between two concepts. Osgood's semantic differential procedure (Osgood, Suci, & Tannenbaum, 1957) required subjects to rate concepts on 10–20 bipolar scales. For example, the concept LADY could be rated on the dimensions of *rough–smooth* and *fair–unfair*, among others. All words of interest would be rated on all of these dimensions. The ratings of a word on these dimensions essentially provides the coordinates of the word in a semantic space, thus making it straightforward to provide a representation for any word for which the set of ratings was collected. Similar words should be located in the same region within the space. This theory motivated the work of Smith, Shoben and Rips (1974) and the development of their semantic-feature model as an alternative to Collins and Quillian's nodal model. This approach provides a rich representation for words. Its drawbacks are that many human judgements are required for each word of interest and that the experimenter must choose the semantic dimensions upon which words are rated (see McRae, de Sa, & Seidenberg, 1997, for a similar, but more recent, normative approach to feature selection).

Semantic features have also been used by those interested in implementing more complete models of comprehension. McClelland and Kawamoto (1986) used a set of distributed representations in their model of thematic role assignment and sentence processing. Words were represented by a set of semantic microfeatures. Nouns, for instance, had features such as HUMAN, SOFTNESS, GENDER and FORM. Verbs had more complex features, such as CAUSE (whether the verb is causal), TOUCH (specifies whether the agent or instrument touches the patient) and NAT_CHNG (what kind of change takes place in the patient). This model was important in that it demonstrated that distributed semantic representations can account for case-role assignment and handle lexical ambiguity. The features selected by the authors served them well; intuitively, the features seem

relevant for thematic-role assignment. Similar approaches to feature designation have frequently been used in the connectionist literature for more basic models of word recognition (Dyer, 1990; Hinton & Shallice, 1991; Plaut & Shallice, 1994). It is less than clear, though, exactly what features one would select for a more general model of semantic representation.

In the last decade, investigators using large-scale corpora have attempted to extract semantic information directly from text. Gallant (1991) has developed a methodology that extracts a distributed set of semantic microfeatures using the context in which a word is found. However, a drawback to his approach is that the features for the core meanings have to be determined by a human judge. The model that we have developed does not rely on any explicit human judgements in determining the dimensions that are used to represent a word and acquires the representations in an unsupervised fashion. Discussion of some important work similar to ours by Landauer and Dumais (1994, in press), Finch and Chater (1992) and Schutze (1992) will be included in the General Discussion, and we will now turn to how the word representations are acquired in HAL.

DEVELOPING SEMANTIC REPRESENTATIONS

Developing a plausible methodology for representing the meaning of a word is central to any serious model of memory or language comprehension. We use a large text corpus of ~300 million words to initially track lexical co-occurrence within a 10-word moving window. From the co-occurrences, we develop a 140,000 dimensional context space (for full implementational details, see Lund & Burgess, 1996; see Landauer & Dumais, in press, for a similar model). We refer to the high-dimensional space as "context" space, since each vector element represents a symbol (usually a word) in the input stream of the text. Each symbol is part of the textual context in the moving window.

Matrix Construction. The basic methodology of the simulation is to develop a matrix of word co-occurrence values for a given vocabulary. This matrix will then be divided into co-occurrence vectors for each word, which can be subjected to analysis for meaningful content. For any analysis of co-occurrence, one must define a window size; that is, the largest number of words that may occur between a pair of words such that the pair may be considered to co-occur. The smallest usable window would be a width of one, corresponding to only immediately adjacent words. At the other end of the spectrum, one may count all words within a logical division of the input text as co-occurring equally (see Landauer & Dumais, 1994, in press; Schvaneveldt, 1990). A very small window may miss constructs spanning several words (lengthy noun phrases, for instance), while large windows risk

introducing large numbers of extraneous co-occurrences. Therefore, we chose a window width of 10 words to preserve locality of reference while minimising the effects of different syntactic constructions. As a further move away from dependence on syntax (or any structuring of the language under consideration other than that given by the division of words), sentence boundaries are ignored.

Within this 10-word window, co-occurrence values are inversely proportional to the number of words separating a specific pair. A word pair separated by a 9-word gap, for instance, would gain a co-occurrence strength of one, while the same pair appearing adjacently would receive an increment of 10. Cognitive plausibility was a constraint, and a 10-word window with decreasing co-occurrence strength seemed a reasonable way to mimic the span of what might be captured in working memory (Gernsbacher, 1990). The product of this procedure is an $N \times N$ matrix, where N is the number of words in the vocabulary being considered. It is this matrix which we will demonstrate contains significant amounts of information that can be used to simulate a variety of cognitive phenomena. A sample matrix is shown in Table 1. This sample matrix models the status of a matrix using only a 5-word moving window for just one sentence, *the horse raced past the barn fell.*

Text Source. The corpus analysed was approximately 300 million words of English text gathered from Usenet. All newsgroups containing English text were included. This source has a number of appealing properties. It was clear that to obtain reliable data across a large vocabulary, a large amount of text would be required. Usenet was attractive in that it could supply indefinitely about 20 million words of text per day. In addition, Usenet is conversationally diverse. Virtually no subject goes undiscussed; this allows the construction of a broadly based co-occurrence dataset. This turns out to be useful when attempting to apply the data to various stimulus sets, for there is little chance of encountering a word not in the model's vocabulary. One goal for HAL was that it would develop its representations from conversational text that was minimally preprocessed, not unlike human-concept acquisition. Unlike formal business reports or specialised dictionaries, which are frequently used as corpora, Usenet text better resembles everyday speech. That the model works with such noisy, conversational input suggests that it can deal robustly with some of the same problems that the human-language comprehender encounters.

Vocabulary. The vocabulary used for the analysis consisted of the 70,000 most frequently occurring symbols within the corpus. About half of these had entries in the standard Unix dictionary; the remaining items included

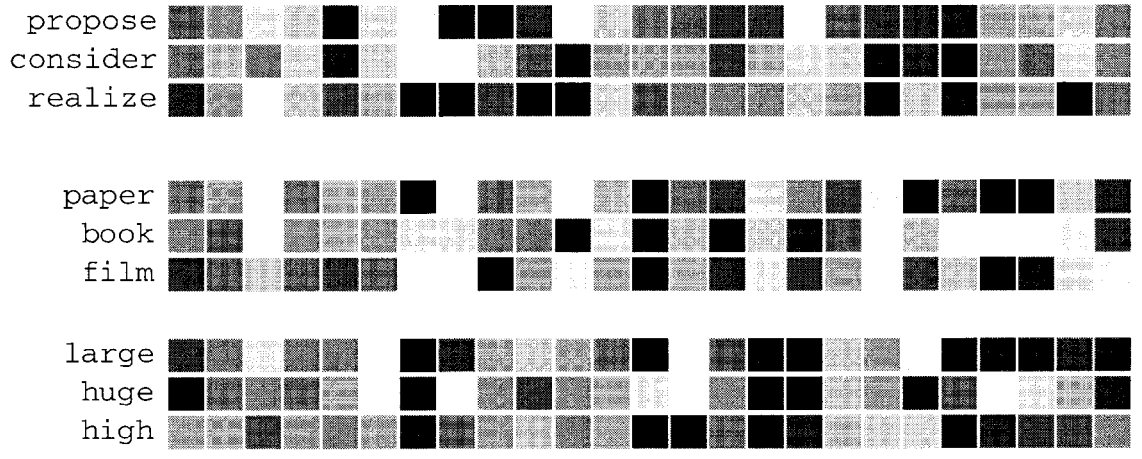

FIG. 1. Sample 25-element word vectors for three sets of verbs, nouns and adjectives. Each vector element has a continuous value (the normalised value from its matrix cell) and is grey-scaled to represent the normalised value with white corresponding to a zero.

TABLE 1
Sample Matrix, Produced by Applying a Five-word Co-occurrence Window to the Sentence ***The horse ran past the barn fell***

	barn	*horse*	*past*	*raced*	*the*
barn		2	4	3	6
fell	5	1	3	2	4
horse					5
past		4		5	3
raced		5			4
the		3	5	4	2

Note: Values in matrix rows represent co-occurrence values for words which preceded the word (row label), while columns represent co-occurrence values for words following the word (column label). Cells containing zeroes have been left empty for visual clarity.

proper names, slang words, nonword symbols and misspellings. These items also presumably carry useful information for concept acquisition.

Data Extraction. The co-occurrence tabulation produces a 70,000 × 70,000 matrix. Each row of this vector represents the degree to which each word in the vocabulary preceded the word corresponding to the row, while each column represents the co-occurrence values for words following the word corresponding to the column. A full co-occurrence vector for a word consists of both the row and the column for that word. The following experiments use groups of these co-occurence vectors.

These vectors (length 140,000) can be viewed as the coordinates of points in a high-dimensional space, with each word occupying one point. Using this representation, differences between two words' co-occurrence vectors can be measured as the distance between the high-dimensional points defined by their vectors.

Vector Properties. As described above, each element of a vector represents a coordinate in high-dimensional space for a word or concept,[1] and a distance metric applied to these vectors presumably corresponds to

[1]Each vector corresponds to one lexical symbol that occurred in the input stream. We will often refer to the vectors as corresponding to words; however, it is important to bear in mind that all symbols were not words, it is just stylistically more convenient to refer to them this way. Vectors of 140,000 element length were used in the simulations reported in this paper. This is not to suggest that we think human memory is best represented in 140,000 dimensional space. At present, we think that 100–200 dimensional space is an accurate estimate of what is required for simulating human memory. For some effects, as few as 10 vector elements can account for a significant amount of variance. In connectionist models, shorter vectors would be preferable due to the training time required for learning in a network. In shortening vectors, it is important

context similarity. The vectors can also be viewed graphically, as shown in Fig. 1. Sample words (propose, film, etc.) are shown with their accompanying 25 element vectors (only 25 of the 140,000 elements are shown for viewing ease). Each vector element has a continuous numeric value (the normalised value from its matrix cell). A grey-scale is used to represent the normalised value, with black corresponding to a zero or minimal value. A word's vector can be seen as a distributed representation (Hinton, McClelland, & Rumelhart, 1986). Each word is represented by a pattern of values distributed over many elements, and any particular vector element can participate in the represenation of any word. The representations gracefully degrade as elements are removed; for example, there is only a small difference in performance between a vector with 140,000 elements and one with 1000 elements. Finally, it can be seen that words representing similar concepts have similar vectors, although this can be subtle at times. (See Lund and Burgess, 1996b, for a full description of the HAL methodology.)

The advantage of representing meaning with vectors such as these is that, since each vector element is a symbol in the input stream (typically another word), all words have as their "features" other words. This translates into the ability to have a vector representation for abstract concepts (e.g. *justice*, *reality*) as easily as one can have a representation for more basic concepts (e.g. *dog*, *book*). This is important, if not absolutely crucial, when developing a memory model that purports to be general in nature.

SIMULATION 1: DEMONSTRATING BASIC CATEGORISATION

So far, only the visual inspection of nine vectors has been offered as evidence that the most informational vector elements can be extracted from the matrix such that they provide a representation that is semantic in natue. In this first experiment, four semantic categories of words[2] are subjected to multidimensional scaling to determine if the interword distances are semantically meaningful.

---

to keep the vector elements that are most "informative". To determine this, the column and row variances for a particular word's vectors to be used in an experiment are computed. The vector elements with the smallest variances can be discarded, since they represent contexts in which the word had little experience. We find that variance drops sharply across the first 100 elements and is very low by the 200th element. Accordingly, to generate a 200 element vector, the 139,800 columns with the lowest variance would be discarded for each word. Empirically, using a long or short vector provides similar results. For the present purposes, it was more computationally straightforward to use the long vector. Further details can be found in Lund and Burgess (1996b).

[2]Word and concept will be treated as equivalent in this paper. We are mindful that it is controversial to do so (Komatsu, 1992), but do not think it detracts from our current theoretical development of the HAL model.

Methods

A number of words that represented four categories (animal types, body parts, cities and geographical locations) were used in an initial analysis of semantic content for the co-occurrence vectors. Vectors were extracted for these words and, treating each vector as a set of coordinates in a high-dimensional Euclidean space, a distance matrix was formed. Our hypothesis was that this distance matrix, representing the interword distances for the chosen set of words, would operate as a similarity matrix. Each element in the similarity matrix represented the distance between two of the chosen words in this high-dimensional space.

Results

This matrix was analysed using a multidimensional scaling algorithm (MDS) which projects points from a high-dimensional space into a lower-dimensional space in a non-linear fashion and attempts to preserve the distances between points as much as possible. The lower-dimensional projection allows for the visualisation of the spatial relationships between the co-occurrence vectors. The two-dimensional MDS solution is shown in Fig. 2. Visual inspection suggests that the four categories of words were differentiated by this procedure. Geographical regions and cities are distinct from the animals and body parts. There is some overlap among the more related categories (locations and cities; animals and body parts). Inferential statistics were computed comparing intragroup distances to intergroup distances. To do this, distances between all combinations of item pairs within a group were calculated and compared to all combinations of the group items and all other items. Animals (denoted by filled stars) were differentiated from the other three groups of items [$F(1,258) = 91.14$, $P < 0.0001$]. Likewise, body parts (denoted by open circles) were different from animals, cities and locations [$F(1,553) = 408.47$, $P < 0.0001$], and the intragroup distances for locations (denoted by open stars) differed from the other three groups [$F(1,304) = 381.45$, $P < 0.0001$]. Intragroup distances for cities (denoted by filled circles) also differed from the other three groups [$F(1,220) = 126.09$, $P < 0.0001$]. However, two pairs of categories did show some degree of overlap, so additional comparisons were conducted. Intragroup distances for animals differed from the intergroup distances to the body parts [$F(1,152) = 28.97$, $P < 0.0001$]. Intragroup distances for cities differed from the intergroup distances to the locations [$F(1,88) = 17.01$, $P < 0.0001$]. Thus the two-dimensional MDS of the 140,000 dimensional space would seem to be an accurate reduction of dimensionality.

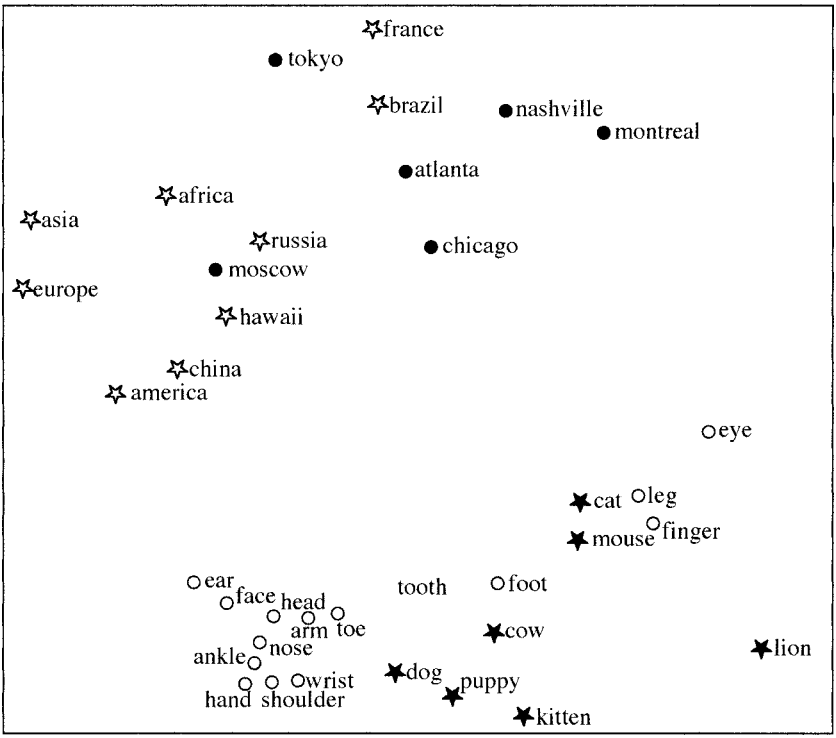

FIG. 2. Two-dimensional multidimensional scaling solution of four groups of words (geographic locations, cities, animals, body parts).

Discussion

Given that words with similar meanings tended to be close to each other in the high-dimensional space, we conclude that these vectors from the co-occurrence matrix carry information that mimics semantic knowledge. Specifically, it would appear that the distances between items reflect some aspect of the semantic relationship of those items. The MDS results presented here are very similar to the results obtained by various semantic memory investigators who employed human ratings using a semantic differential procedure (Osgood et al., 1957). There is a long history in cognitive psychology of using rating scales with some predetermined set of adjectives as bipolar anchors which essentially determine a set of "semantic features" (Smith et al., 1974). HAL does not use a set of "semantic" features; vector elements correspond to symbols that actually occurred in the learning environment. Certain of these elements contribute more heavily to the results obtained (those with the highest variance). We will argue later for the

theoretical importance of these contextual elements. Using contextually based representations such as these allows one to avoid the need to commit to a set of semantic features that can be difficult to justify on theoretical grounds.

From a pragmatic viewpoint, it is difficult to obtain a sufficient number of human observations to develop an adequate set of representations. Using the minimum guidelines offered by Osgood et al. (1957), obtaining a 10-dimensional set of coordinates for 100 words would require 50 raters. This translates into 50,000 individual human judgements. To have a 20-dimensional set of representations for 3000 words (similar to the simulation without a semantic system by Seidenberg & McClelland, 1989) using a more conservative 100 human raters would require 6 million individual judgements. Although human normative data have been used in computational models (Burgess & Lund, 1994; McRae et al., 1997), there is considerable overhead associated with this approach. This is most likely the reason behind the two most frequently used approaches to modelling semantic representations: binary vectors that do not correspond to actual words (Kawamoto, 1993; Masson, 1995; Moss, Hare, Day, & Tyler, 1994), or simply constructing a set of vector features that presumably will represent a word's meaning (Hinton & Shallice, 1991; McClelland & Kawamoto, 1986; Plaut & Shallice, 1994).

The results reflected in the MDS of the items in this simulation suggest that the distances between words might roughly correspond to reaction-time latencies from single-word priming experiments. However, there has been a longstanding debate over whether word priming may, in fact, be carried by associative relationships rather than semantic similarity (Lupker, 1984; Mandler, 1962; Shelton & Martin, 1992). Elsewhere we have presented evidence that HAL's vectors are more semantic than associative in nature by looking at the correspondence between the priming effects of related and unrelated sets of item vectors and the results of experiments with human subjects (Lund et al., 1995, 1996). In addition, we have found HAL's vectors to accurately reflect human judgements made on Toglia and Battig's (1978) semantic word norms and the typicality effect (Lund & Burgess, 1994). The results of this simulation and of the experiments just cited all converge on the notion that the information carried by the vectors is more semantic than associative in nature.

Earlier, we suggested that the semantic characteristics of the vectors most likely hinge on the fact that similar words are used in similar contexts. As the 10-word window moves along the text, the matrix develops in a way that can be sensitive to the contexts in which the words occur (Lund & Burgess, 1996b). This bootstrapping procedure capitalises on the contextuality inherent in the input stream to create a set of vectors that carry this semantic, categorical knowledge reflected in the results of simulation 1. Of course, the

notion that the contexts that a word occurs in results in knowledge about the word is an idea with a long history in psycholinguistics. Early models of memory recognised that one basic component of memory organisation was grammatical categorisation (Deese, 1965). Most associative responses in a word association task are of the same grammatical class (Ervin, 1963). Deese (1965, ch. 5) demonstrated that a factor analysis of word associations would produce a cluster of different grammatical classes of verb forms and types of pronouns. More recently, Slator et al. (1990) used a machine-readable dictionary analysis to categorise types of prepositions. The dictionary they used (LDOCE) includes a variety of information such as synonyms and short phrases, as well as short sample texts illustrating how a word can be used. LDOCE, however, provides a considerable amount of information to a model up front so that it may learn grammatical categorisation. Although this is a valuable way to jumpstart grammatical categorisation, it is not as satisfying as an approach that can provide the necessary representations to do this strictly from conversational text.

SIMULATION 2: SIMPLE GRAMMATICAL CATEGORISATION

Ervin (1961, 1963) and others (see Nelson, 1977, for a review) have presented considerable evidence that a child's experience with context results in both semantic and grammatical categorical knowledge. More recently, Finch and Chater (1992) demonstrated that words can be categorised according to grammatical class using a learning procedure very similar to ours. Although there is evidence that HAL's word vectors carry semantic information, Finch and Chater's results suggest that HAL's vectors may carry a broader range of categorical knowledge. One potentially important distinction between the two simulations is that Finch and Chater used a 2-word moving window, whereas we used a 10-word window. The possibility exists that the larger encoding window may not be sensitive to the more local syntactic dependencies. The goal for this experiment was to see if HAL's vector representations, generated using our 10-word window, can provide the required information to distinguish basic grammatical categories similar to that found by Finch and Chater (1992).

Method

The first analysis used vectors for items for four grammatical categories: nouns (10), verbs (10), determiners (9) and prepositions (6). The vector representations were extracted for these verbs and analysed using a MDS algorithm which projects the high-dimensional space into a lower-dimensional space while attempting to preserve the interpoint distance.

Results

The MDS solution for these four grammatical classes is shown in Fig. 3. Visual inspection suggests that the scaling procedure was able to capitalise on the information carried in the vector representations. Verbs (denoted by a triangle), for the most part, are found clustering on the right, whereas nouns (denoted by a circle) cluster on the left. Three exceptions of verbs that did not cluster with other verbs (*study*, *sketch* and *request*) are verb–noun ambiguities and clustered more closely with the nouns. The function words are all clustered between the nouns and verbs and seem to have their own independent organisation. The determiners (denoted by squares) show some substructural organisation; the plural possessive pronouns, *their* and *our*, are close, as are the singular possessive pronouns, *my*, *his* and *her*. The possessive pronoun, *your*, is ambiguous in that it can be either singular or plural, and is found between the singular and plural pronouns. The articles, *the*, *a* and *an*, were also close in proximity, with the indefinite articles closer to each other than to the definite article. Likewise, the prepositions (denoted by asterisks) show similar organisation with some correspondence to semantic similarity. At the top of Fig. 3 are *above* and *below*, and in the centre are *in*, *on*, *around* and *by*. This two-dimensional representation of this high-dimensional space is less visually compelling, however, than the results shown with the semantic categories. For example, the prepositions *above* and *below* are closer to the nouns *bed* and *roof* than they are to the other

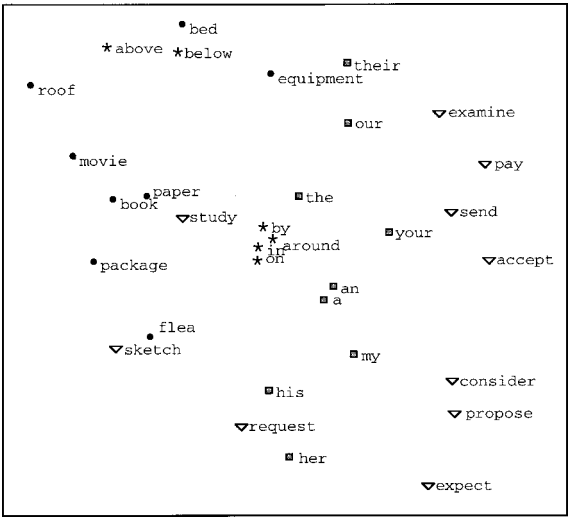

FIG. 3. Multidimensional scaling solution for verbs (inverted triangles), nouns (circles), prepositions (asterisks) and determiners (squares).

prepositions. An analysis of variance comparing each grammatical category's intragroup distances to all other intergroup distances suggests a clear separation in the high-dimensional space. Determiners are differentiated from the other three grammatical categories [$F(1,295) = 238.17$, $P < 0.0001$]. Nouns differ from the other categories [$F(1,262) = 20.36$, $P < 0.0001$], as do prepositions [$F(1,196) = 69.97$, $P < 0.0001$] and verbs [$F(1,328) = 28.75$, $P < 0.0001$]. Thus one can be reasonably confident that the high-dimensional resolution of these basic grammatical categories represents salient information carried in the word vector, and is similar to the results obtained by Finch and Chater (1992).

Discussion

These results compare favourably with the various analyses of grammatical classes by Deese (1965, ch. 5), who conducted a factor analysis of responses found in human association norms, as well as the more recent results of Finch and Chater (1992), whose vector representation procedure was similar to ours. It was important to determine that the model with the 10-word window that produced the apparent semantic categorisation in simulation 1 would also produce the grammatical categorisation found by Finch and Chater. The results from these simulations suggest that the vector representations carry important grammatical-class information sufficient to allow one to distinguish between major grammatical entities (nouns, prepositions, determiners and verbs).

It would appear plausible that the aspects of grammaticality carried in HAL's vector representations could be used by the parser as convergent information in tagging parts of speech, as well as carrying more conventional semantic information. These results are consistent with our earlier idea (Lund et al., 1995) that the categorical information carried in the vector representations emerge, in some abstract way, due to the substitutability of words in contexts. To any extent that this is true, both grammatical-categorical knowledge and semantic-categorical knowledge would seem to be an integral part of this substitutability process, as well as being implicit in a word's vector representation.

SIMULATION 3: MORPHOLOGICAL AMBIGUITY OF VERBS

The results from simulation 2 suggest that HAL's representations carry sufficient information to distinguish between discrete grammatical classes. In this simulation, we wanted to determine if a more subtle grammatical categorisation is possible. We tested this using a set of morphologically ambiguous verbs (*examined*) that were subjected to a MDS procedure, together with verbs that are unambiguously past-tense or past-participle

(*threw* and *grown*, respectively). Comparing ambiguous verbs to unambiguous verbs is of particular interest, since the final simulation in this paper will focus on the contextual effects of nouns and morphologically ambiguous verbs that have been used in on-line syntactic processing experiments. As a precursor to that, it would be important to demonstrate that contextual information relevant to morphological ambiguity in verbs is carried in a word's representation.

Method

Three types of verbs were used: morphologically ambiguous verbs ($n = 26$) (past-tense or past-participle), morphologically unambiguous past-tense verbs ($n = 9$) and morphologially unambiguous past-participle verbs ($n = 9$). The ambiguous verbs were the verbs used in the study of Burgess and Tanenhaus (1996). The unambiguous past-participle verbs were items used in a previous control experiment by these authors. The unambiguous past-tense verbs were quasi-randomly selected from a list of unambiguous verbs listed in Irmscher (1972). The only constraint on the selection was that verbs that were noun–verb ambiguities (e.g. *saw*) were not used. The vector representations were extracted for these verbs and analysed using a MDS algorithm which projects the high-dimensional space into a lower-dimensional space while attempting to preserve the interpoint distance.

Results

The results of the second MDS with the morphologically ambiguous and unambiguous verbs are shown in Fig. 4. The past-tense verbs (denoted by open circles) can be seen at the bottom right of the figure, whereas the past-participle verbs (denoted by solid dark circles), at the top right of the figure, are clearly separated from the past-tense verbs. Distances were calculated between all possible items pairings allowing for intergroup and intragroup comparisons among the three types of verbs (past-tense, past-participle, ambiguous). Intragroup distances for the past-tense verbs differed from the intergroup distances of the past-tense items to the other two verb types [$F(1,292) = 82.50$, $P < 0.0001$]. Similarly, the intragroup distances for the past-participle verbs differed from the intergroup distances of the past-participle items to the other two verb types [$F(1,418) = 114.17$, $P < 0.0001$]. The exception to the close clustering of the past-participle verbs is *seen*, which can be found with the past-tense verbs. Even so, the two unambiguous verb types differ from each other and from the ambiguous verbs in the high-dimensional space. The ambiguous verbs (denoted by solid, grey-scaled circles) are dispersed throughout the high-dimensional space, as one would expect with verbs that are ambiguous, and their intragroup distances do not differ from the intergroup distances

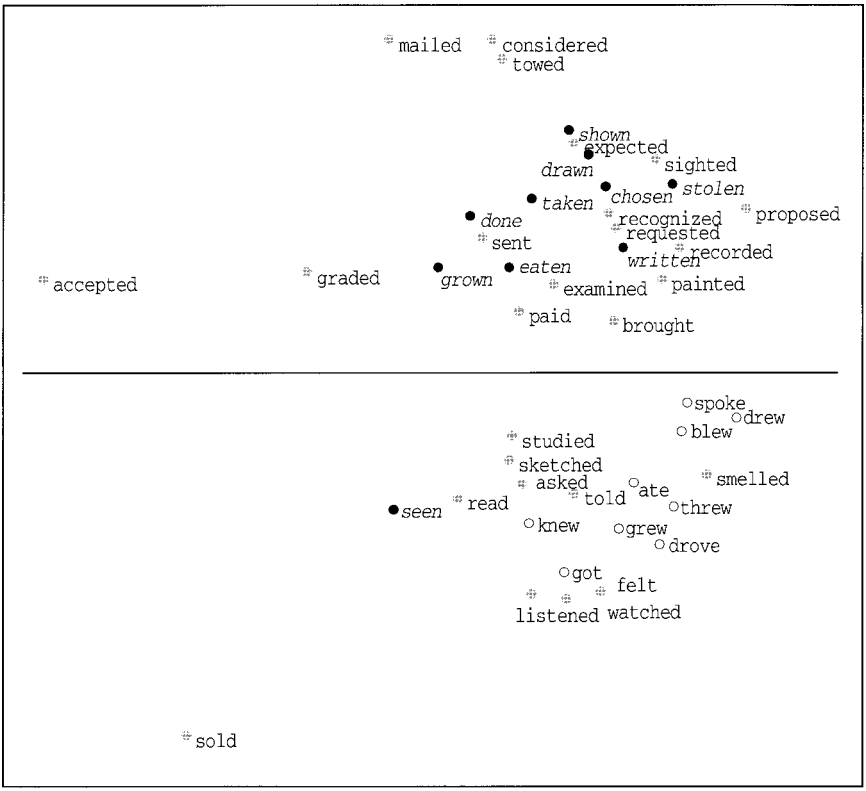

FIG. 4. Multidimensional scaling solution for morphologically ambiguous and unambiguous verbs. Past-tense verbs (open circles), past-participle verbs (solid dark circles) and morphologically ambiguous (solid, grey-scaled circles).

[$F(1,1048) = 0.00$, $P > 0.99$]. As ambiguous verbs, they share space with either type of unambiguous verb. From an earlier simulation experiment (Burgess & Lund, 1994), we had available the relative frequencies with which these ambiguous verbs were used in past-tense and past-participle structures (all sentences using these verbs were from the Brown corpus). One would expect a tendency for the ambiguous verbs in the top half of the figure (clustering near the past-participle verbs) to be used more frequently in past-participle constructions. Likewise, one would expect a tendency for the ambiguous verbs in the bottom half of the figure (clustering near the past-tense verbs) to be used more frequently in past-tense constructions. We conducted an analysis to determine if this was true. The ambiguous verbs that clustered around the past-participle verbs (the ambiguous verbs in the top half) had a mean past-tense bias of 70.4% (range = 33–90%); the

ambiguous verbs that clustered around the past-tense verbs (the ambiguous verbs in the bottom half) had a mean past-tense bias of 87.1% (range = 50–98%). To determine if this was a reliable difference, each ambiguous verb was coded as clustering with either the past-participle or past-tense verbs (top half and bottom half of Fig. 4, respectively). This code was used as a dependent variable in a regression analysis to determine if the actual past-tense bias percentage that was calculated from the corpus analysis would accurately predict the cluster location (corresponding to the top half, bottom half code) of the ambiguous verbs. The past-tense bias did predict cluster location, accounting for 21.6% of the variance [$F(1,23) = 6.36$, $P < 0.02$].

Discussion

The two unambiguous verb forms separate nicely in the high-dimensional space. The verbs that are ambiguous in morphological form are located throughout the space occupied by the unambiguous forms. This occurs as a function of frequency of use in past-tense or past-participle sentential contexts as derived from a corpus analysis.

The results from this simulation suggest that the vector representations carry important grammatical-class information which captures the more subtle aspects within a major grammatical class (morphological forms). Moreover, the nature of this pattern of results fits with our notion of morphological verb ambiguity (Burgess & Hollbach, 1988; Burgess & Lund, 1994). Presumably, the vector's sensitivity to the bias of the morphological ambiguity (past-tense, past-participle bias) is a function of the nature of the contexts in which the verbs are used. Past-tense and past-participle verbs vary distributionally with the semantic characteristics of the nouns that serve as their subjects and objects. This variation has been exploited in psycholinguistic research investigating semantic constraints during syntactic processing. An analysis of stimuli from a set of these experiments will be the focus of the last simulation.

SIMULATION 4: SINGLE-WORD SEMANTIC CONTEXTS AND PARSING

The degree to which lexical/semantic information can assist syntactic processing during on-line sentence comprehension is determined to a large extent by the architecture of the language processor and the constraints offered by the constituents in a sentence. In this experiment, we demonstrate that the context distances which can be derived using the HAL model correspond to the constraint of the semantic contexts used in three different experiments.

Syntactically motivated theories, most notably the garden path theory of Frazier and colleagues (Ferreira & Clifton, 1986; Frazier, 1978; Rayner et al., 1983), describe a set of deterministic parsing strategies that are insensitive to contextual information, relying solely on structural cues. One of these strategies, minimal attachment, proposes that words are attached to a phrase marker using the fewest possible syntactic nodes. As a result, the parser will initially build a past-tense construction (see Fig. 5a) rather than the ultimately correct past-participle analysis (see Fig. 5b).

A series of early experiments had demonstrated that various discourse and semantic contexts were ineffective in reducing the processing load associated with the garden-pathing phenomenon (Ferreira & Clifton, 1986; Rayner et al., 1983). Ferreira and Clifton's (1986) study is probably one of the best illustrations of such an experiment. They reasoned that Rayner and co-workers' (1983) attempt may have failed due to weak pragmatic constraints expressed in the NP + V combinations. As a stronger manipulation, they varied the animacy of the sentential subject such that it was a virtually anomalous agent for the subsequent verb, so that, if the parser could use this semantic information, it would eliminate the garden path in sentence (1b) but not in (1a):

1a. The defendant examined by the lawyer turned out to be unreliable.
1b. The evidence examined by the lawyer turned out to be unreliable.

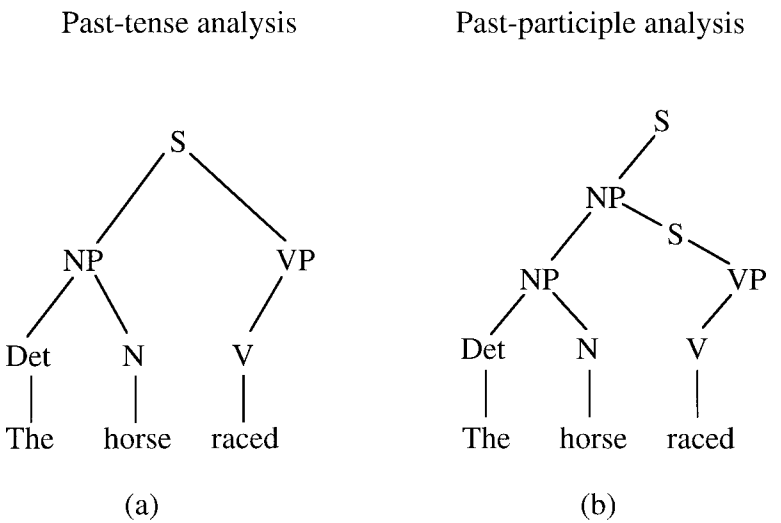

FIG. 5. Abbreviated phrase structure trees for the past-tense (a) and the past-participle (b) analyses of *the horse raced* . . .

In their experiments, the animacy manipulation had no effect on first-pass reading times in the by-phrase, providing further empirical support that the minimal attachment preference results in a garden path regardless of the semantic plausibility of the noun. Considerable evidence exists suggesting that the parser is initially insensitive to pragmatic and semantic constraints and that the semantic information is used only to assist in reparsing a sentence once misassignments have been made (see Clifton & Ferreira, 1989, for a review).

Research in lexical ambiguity has motivated the development of constraint-based parsing models (Burgess & Hollbach, 1988; Burgess & Lund, 1994; MacDonald et al., 1994; Tanenhaus & Carlson, 1989). These models allow more flexibility for communication between language subsystems and suggest a larger role for lexical/semantic representations in that they can provide guidance for the necessary lexical, thematic and syntactic commitments that are required during on-line language comprehension. Semantics in these models plays a crucial role in the parsing operation. Real-world knowledge and semantic information (see MacDonald et al., 1994; Tanenhaus & Carlson, 1989) can provide immediate feedback to the parsing operation such that cognitive-processing loads (i.e. garden paths) that are obligatory on a minimal attachment view can be eliminated. There is increasing empirical evidence to support such models (Burgess & Tanenhaus, 1996; Burgess & Lund, 1994; MacDonald et al., 1994; Trueswell, Tanenhaus, & Kello, 1993; see MacDonald et al., 1994, for a review). Two of these studies will be discussed here, since their stimuli and the stimuli used in Ferreira and Clifton's (1986) study will be included in our simulation.

Burgess and Tanenhaus (1996; Burgess, 1991) conducted an experiment along the lines of Ferreira and Clifton (1986, experiment 1) using sentences similar to (1a) and (1b). However, Burgess and Tanenhaus had noted that a number of the sentences used by Ferreira and Clifton were not as constraining as might be desired (see Burgess, 1991, for details). For example, sentences with inanimate subjects, such as (2a) and (2b), presumably are biased for the past-participle interpretation because the subject noun will not make a plausible agent:

2a. The car towed from the parking lot was parked illegally.
2b. The ship sighted by the lookout probably brought bad news.
2c. The car towed the boat to the beach.
2d. The ship sighted the survivors.

It is possible, however, to continue a number of these sentences in a past-tense fashion (see 2c and 2d) which compromises the past-participle bias that was intended. Burgess and Tanenhaus conducted a norming study to construct a set of sentences that would have a more compelling

past-participle constraint, and thus provide a stronger test of the semantic constraint. These items, with more constraining agent–verb relationships, showed that the processing load which Ferreira and Clifton associated with the garden-path effect could be eliminated. MacDonald (1994) conducted a similar experiment that more precisely manipulated several of these stimulus constraints. Using reduced-relative sentences (like these previous investigators), she also found that the agent–verb constraint, along with argument structure frequencies and post-ambiguity constraints, helped govern whether or not readers experienced a processing load at the point of disambiguation. The results of the studies of Burgess and Tanenhaus and of MacDonald demonstrate that with strongly enough constrained past-participle sentences, the garden path predicted by minimal attachment can be avoided. This suggests that semantic information can play a role in the initial syntactic processing during sentence comprehension.

The three syntactic processing experiments briefly reviewed here have in common a concern about the nature of the semantic/pragmatic relationships inherent in the sentence. Specifically, these three experiments all investigated the past-participle constructions in which the semantic context was posited as providing guidance to the parser via the relationship between the NP and the morphologically ambiguous verb. These experiments differed with respect to whether or not the garden path associated with the syntactic misassignment occurred. Of concern for this simulation is the extent to which the context distances between the noun–verb combinations used in these three experiments will correspond to the empirical results obtained. During the vector acquisition process, the HAL model is exposed to ~300 million words of text. As a result, verbs encounter a rich array of contexts in which they are used, and this experience is captured in the vector representations of the words. It follows, then, that the vector representations, which encode context, should reflect the probabilistic constraints of the language and that these constraints may have some correspondence to the NP–verb constraints being discussed here. Closer context distances should be an indicator of more constrained noun–verb relationships. Thus close context distances should be associated with the experiments that demonstrated the effects of context (i.e. Burgess & Tanenhaus, 1996; MacDonald, 1994), whereas the experiment that did not show a contextual effect should have noun–verb pairs with longer context distances (i.e. Ferreira & Clifton, 1986).

Method

Stimuli. Noun–verb pairs from the sentential stimuli were used from the three studies. Animate and inanimate stimuli from Ferreira and Clifton (FC) (e.g. *baby–felt*, *skin–felt*) yielded 32 noun–verb pairs (see their appendix 1).

Likewise, the animate and inanimate stimuli from Burgess and Tanenhaus (BT) (e.g. *man–paid*, *ransom–paid*) yielded 32 noun–verb pairs (see their appendix 1). Finally, the animate and inanimate stimuli from MacDonald (MM) (e.g. *spy–concealed*, *microfilm–concealed*) yielded 32 noun–verb pairs (see her appendix 3). All experiments have in common that the sentences were biased towards either a past-tense (animate) or a past-participle (inanimate) interpretation. The constraint to be evaluated by this experiment was only between the noun–verb pairs of these three stimuli sets.

Procedure. The design of the experiment was a 2 × 2 mixed factorial; both factors were between items. Bias had two levels corresponding to whether the noun–verb pairs were PTC or PPC. Study had two levels corresponding to the experiment that did not obtain an effect of context (FC) and the experiments that did show an effect of context (BT and MM). All analyses were calculated on item manipulations. Recall from the simulation methodology that the matrix develops from tracking co-occurrences using a 10-word window that slides along the text one word at a time. Earlier pilot work had demonstrated that optimal results when investigating parsing constraints are obtained with window sizes of six or eight. Therefore, for this experiment, we extracted two sets of vectors: one from a matrix using a window size of six and the other using a window size of eight. As a result, each set of stimuli from each of the three studies to be simulated has two sets of vector representations. Context distances were computed for all noun–verb pairs and are presented in RCUs (Riverside Context Units[3]), which is the distance metric used in the HAL model.

Results

The results are shown in Fig. 6. The experimental design was a 2 (bias: PTC or PPC) × 2 (study set: those showing context effects, BT & MM, and the one that did not show the effect of noun context, FC). A complete vector representation consists of 70,000 elements from the rows of the matrix and 70,000 elements from the columns of the matrix. Rows encode the information in the *n*-word moving window that occurs before the target word; columns encode the information in the *n*-word moving window that occurs after the target word. In this simulation, rows and columns were separated, since subject position nouns tend to occur before the verb in English, and thus row information might be more informative in this

[3]We refer to the distance metric used in the HAL model as Riverside Context Units (RCUs). This is an arbitrary, but normalised Euclidian distance measure (see Lund & Burgess, 1996b, for further description). To normalise a vector, we first compute its magnitude (sum the squares of its elements, take the square root of that, and divide by the number of elements). We then divide the magnitude by 666.0, and divide each vector element by the resulting number.

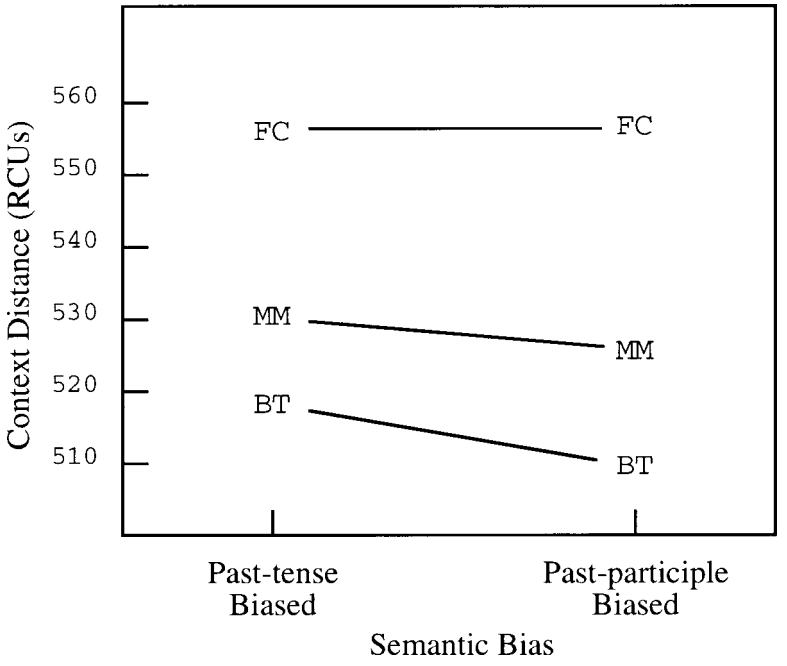

FIG. 6. Context distance as a function of simulated experiment and semantic bias.

simulation.[4] For rows, there was a main effect of study [$F(1,248) = 4.45$, $P = 0.036$] but no main effect of bias ($F < 1$), nor was the interaction reliable ($F < 1$). For columns, there was a marginal main effect of study [$F(1,246) = 296$, $P = 0.053$] but no main effect of bias ($F < 1$), nor was the interaction reliable ($F < 1$). The two studies that showed the effect of the semantic context of the noun on parsing decisions (BT, MM) showed shorter noun–verb semantic distances than the study that did not obtain the context effect (FC). The difference between the row and column analysis was negligible.

Discussion

The more recent class of constraint satisfaction models of syntactic processing allow a role for contextual constraints that become activated during the word recognition process (Burgess & Lund, 1994; MacDonald et al., 1994; Tanenhaus & Carlson, 1989; Trueswell, Tanenhaus, & Garnsey, 1994). Modelling parsing constraints, however, offers a more serious challenge for a memory model. Generally, different syntactic alternatives will give rise to very different semantic or thematic constraints in the

[4]We want to thank Ken McRae for suggesting the comparison of the row and column results.

sentence. Discourse context may be required to confirm partial analyses. The stimuli from the three experiments which we used in this simulation have in common that the first noun is conceivably an agent or a patient in the initial analysis of the sentence. Being an agent or a patient entails a very different semantic or thematic relationship to the verb. The relationship between the verb and its agents and patients implicates certain syntagmatic constraints in language, and the results of this simulation experiment suggest that the vectors carry at least some of that type of information.

While one might expect the context distance to correspond to the within-experiment bias variable (PTB *vs* PPB), previous experience suggests that this will not occur (Burgess & Lund, 1994). For example, in an experiment that manipulates the animacy of the noun, both the animate and inanimate nouns constrain the verb for a particular morphological form, but do this in different ways which correspond to the plausibility of their being an agent or a patient. The degree to which this succeeds can relate to context distance and seems to vary across experiments, not within. That is, experiments that showed a context effect probably had nouns that were better agents for the verb as well as more appropriate inanimate patients. HAL's context vectors do not carry any explicit information about whether a word is a good or bad agent. Other models have explicitly coded agency information as part of their word-vector representations (McClelland & Kawamoto, 1986; St. John & McClelland, 1990). A model like HAL does not have agency information encoded in its vectors in any direct way (if it did, perhaps the row/column manipulation would have been more pronounced). HAL's representations are encodings of the history of the contexts in which a word has appeared. The distance between words in the high-dimensional space reflects this learning history, and the distance metric appears to predict whether the contexts were sufficiently constraining in the three parsing experiments investigated in this simulation. These vector representations are far from perfect in their predictions, however. Pearlmutter and MacDonald (1995) conducted an experiment in which sentences with morphologically ambiguous verbs (e.g. *cooked* in 3a) were contrasted with sentences with unambiguous, intransitive verbs (e.g. *bubbled* in 3b). In this case, a context distance analysis of the stimuli show that the more constraining items for their particular sentences (e.g. *soup bubbled*) have longer context distances than the less constraining items (e.g. *soup cooked*). In HAL, *soup* and *cooked* are more contextually related than are *soup* and *bubbled*. This makes intuitive sense, but it makes clear that these word vectors, as HAL is presently implemented, will not be able to capture the full constraints that are present in a sentence.

3a. The soup cooked in the pot but was not ready to eat.
3b. The soup bubbled in the pot but was not ready to eat.

Overall, the results from this experiment demonstrate that the contextual constraints that are important in determining the occurrence of a processing load during syntactic processing can be modelled using the vector representations generated by the HAL model. Although this experiment did not simulate syntactic processing *per se*, these results also lend support to the notion that contextual constraints are used during the parsing process because of the correspondence of the simulation results to this set of three psycholinguistic experiments.

GENERAL DISCUSSION

In this paper, we have attempted to demonstrate that HAL's vector representations can go a long way towards forming a useful tool for modelling contextual effects in syntactic processing. Having an adequate representational model for the meaning of a word has become crucial, as the focus on parsing constraints moves in the direction of a richer semantic system. In the first simulation we demonstrated that the representations were able to reliably differentiate four semantic categories. However, categorisability extends beyond "semantic" categorisation, and in the second simulation we showed that the vector representations carry information sufficient to categorise distinct grammatical forms (nouns, verbs, prepositions and determiners). The vectors would seem to carry a broad range of categorical information corresponding to what is typically considered to be semantic and grammatical. The third simulation with the unambiguous and ambiguous verbs demonstrates some important characteristics for representations if they are to be viable in modelling parsing constraints. HAL's vector representations for unambiguous past-tense and past-participle verbs clearly differentiated these two morphological forms. Morphologically ambiguous verbs were distributed throughout the high-dimensional space. In addition, their relative location in the space (close to the past-tense or past-participle) was related to the relative frequency of use as a past-tense or a past-participle. Models of language comprehension that posit a strong role of the lexical/semantic system require that a word's representation be rich in content. Sentence processing constraints are varied, but certainly include semantic and grammatical information and some ability to model lexical/morphological ambiguity. The results of the first three simulations suggest that HAL's vector representations have encoded information that allows for at least some basic processing along these constraint dimensions. The final simulation involved the specific constraint that a noun context has on a morphologically ambiguous verb. This type of reduced-relative structural ambiguity has a long history in psycholinguistics, since it has been used to

test different models of syntactic processing. The stimuli (the noun–verb pairs) from three sets of empirical results were used to simulate the constraint of these noun–verb pairs and to determine if the original results corresponded to the context distances in the high-dimensional space. We found that the stimuli from the experiment that showed no effect of noun context on syntactic processing had longer context distances than the stimuli from the two experiments that did find the context effect. Closer context distances correspond to experiments that show context effects (at least with the three we used). We also indicated how this was not a perfect correspondence. Recall Pearlmutter and MacDonald's (1995) parsing constraint experiment that used intransitive verbs (e.g. *soup–bubbled*) and ambiguous transitive items (e.g. *soup–cooked*). HAL's context distances make the wrong prediction here (since *soup* and *cooked* are generally more contextually related). Clearly, there is linguistically relevant information that does not get captured in these word representations, possibly due to the impoverished input stream (compared to all the sensory input a human has available), or it may be that certain constraints simply cannot be captured in a word's representation. Regardless, though, it seems clear that the representations carry a considerable range of information, much of it potentially useful to a language comprehension system.

The Question of Representational Modularity

Whether or not the syntactic processor can use contextual information to guide its parsing decision has been a controversial issue; the question itself presupposes a parsing mechanism. Recent theories of parsing have been driven by lexical/semantic models of word recognition. The notion of a two-stage parser, where a syntactic structure is built without initial recourse to the available semantics, continues to be a dominant theory in psycholinguistics (Clifton & Ferreira, 1989; Frazier & Clifton, 1996). More recent models of syntactic processing have relied increasingly on the richness of the lexical/semantic system to provide the various semantic, thematic and local co-occurrence information required to correctly assign meaning to word order (Burgess & Lund, 1994; MacDonald et al., 1994; Tanenhaus & Carlson, 1989). Basic constraint satisfaction models are free to use a broad range of information and further acknowledge that these different sources of information vary in their relative contribution to the sentence comprehension process. The evidence that supports a constraint-satisfaction approach calls into question any strict notion of modularity of processing. Recent results suggest that the language processor is not modular, and that whether or not modular performance is observed, is a function of a variety of constraints that may or may not be available.

A parallel issue exists with respect to modularity of representations. Most theories of language comprehension assume that different forms of representations (e.g. syntactic, grammatical, lexical and semantic) are linguistically distinct, regardless of their position on processing modularity (Burgess & Hollbach, 1988; Burgess & Lund, 1994; Frazier, 1978; Frazier & Fodor, 1978; MacDonald et al., 1994; Tanenhaus & Carlson, 1989). Connectionist word recognition models have tended to blur this distinction by consolidating the learning from different representational sources into a single layer of hidden units (Elman, 1990; Seidenberg & McClelland, 1989). HAL's vector acquisition process simply accumulates a word's representation from the word's surrounding context. Each vector element for a particular word corresponds to a symbol (usually another word) in the input stream that was part of the contextual history for that particular word. The word's representation, then, corresponds to the complete contextual learning history that is a function of the word's context, the frequency of co-occurring symbols, and the relative weight in the moving window. The simulations reported in this paper, as well as our previous work with semantic priming, word association norms (Lund et al., 1996) and other syntactic effects (Burgess, Livesay, & Lund, 1996), suggest that HAL's representations carry a broad range of information that accounts for a variety of cognitive phenomena. This generality of HAL's representations suggests that it is possible to encode many "types" of information into a single representation and that this can be contextually driven.

With the increased reliance on contextual factors and their influence in syntactic processing, the need for a representational theory is vital. We propose that the vector representations that are acquired by the HAL model can provide such a resource. These vector representations are a product of considerable language experience (~300 million words of text in these simulations) that reflect the use of words in a highly diverse set of conversational contexts. The model does not presuppose any primitive or defining semantic features, and does not require an experimenter to commit to a particular type or set of features. Rather, the model uses as "features" (i.e. the vector elements) the other words (and symbols) that are used in language. That is, a word is defined by its use in a wide range of contexts.

Global or Local Co-occurrence?

As the lexical/semantic system increases in importance to our thinking about the language comprehension system in general, computational linguists have become increasingly interested in lexical co-occurrence as a mechanism for retrieval (Saint-Dizier & Viegas, 1995; Spence & Owens, 1990). We define local co-occurrence as the statistical likelihood of a

particular word occurring in another's presence over some relatively short window. These local statistics have been useful in developing approaches to resolving syntactic ambiguity (Hindle & Rooth, 1993), in creating lexically oriented information-retrieval systems (Smadja, 1993) and in determining parts of speech (Merialdo, 1994). Juliano and Tanenhaus (1993) have shown that local co-occurrence rates of the word *that* in different syntactic contexts can predict the extent that a reader will make syntactic misassignments during reading. Burgess and Lund (1994) used as one constraint in their connectionist parsing model the probability of the disambiguating preposition *by* occurring after the morphologically ambiguous verb. Word co-occurrence may be an important source of information that a language user can use.

It is important, however, not to confuse local co-occurrence with global co-occurrence. A vector representation generated by the HAL model is a measure of global co-occurrence. Each element in a vector corresponds to a symbol (usually a word) in the input stream. Although the small window that moves along the text enters local co-occurrences into the data matrix, it is the large, sparse data matrix and the vectors that can be extracted from it that form a measure of how a word was used in a broad range of contexts. At a theoretical level, we argue that it is important not to think of these word vectors as simply the concatenation of the local co-occurrences (although it is that). It is the pattern of co-occurrences that is crucial to forming the meaning of a word in HAL. The word vector is, essentially, a contextuality vector (a measure of global co-occurrence). Similar words get used in similar contexts. Using a knowledge acquisition system that records (learns) how words were used in such a large number of diverse contexts, a representation is formed that is the result of considerable language experience. As a result, the vector representation embodies many rich semantic and grammatical aspects of the word that can be useful in modelling cognitive and psycholinguistic phenomena. Landauer and Dumais (in press) argue that their global co-occurrence model essentially exploits these local co-occurrences to induce more general knowledge representations. Their latent semantic analysis (LSA) model, Finch and Chater's (1992) bootstrapping model and our HAL model have much in common, and we agree with Landauer and Dumais' assertion that high-dimensional space models "produce sufficient enhancement of knowledge to bridge the gap between the information available in local contiguity and what people know after large amounts of experience" (p. 2).

Further Representational Issues

Symbol-grounding Problem. Glenberg (in press) raises two issues that he claims are serious problems for most memory models, the first of which is

the symbol-grounding problem. The representations in a memory model do not have any extension to the real world; that is, lexical items cannot be understood with respect to just other lexical items. There also has to be a grounding of the representation of the lexical item to its physical reality in the environment (cf. Cummins, 1996). A model that represents a concept by a vector of arbitrary binary features or by some set of intuitively reasonable, but contrived, set of semantic features does not have a clear mapping onto the environment that it supposes to represent. HAL takes a very different approach to this problem. In HAL, each vector element is a coordinate in high-dimensional space for a word. What is important to realise about each vector element is that the element *is* a direct extension to the learning environment. A word's vector element represents the weighted (by frequency) value of the relationship between the part of the environment represented by that element and the word's meaning. The word's meaning is comprised of the complete vector. Symbol grounding is typically not considered a problem for abstract concepts. Abstract representations, if memory models have them, have no grounding in the environment. Again, though, HAL is different in this regard. An advantage to the representational methodology used in HAL is that abstract representations are encoded in the same way as more concrete words. The language environment (i.e. the incoming symbol stream) that HAL uses as input is special in this way. That is, abstract concepts are, in a sense, grounded.

The second problem faced by models that develop "meaningless" internal representations is that the variety of input that a human can experience does not get encoded and, therefore, the memory representation is inevitably impoverished. With the current implementation of HAL, this is certainly a limitation. The learning experience is limited to a corpus of text. This raises an important but currently unanswerable question: Are the limitations in HAL's representations due to the impoverished input or will higher-level representations be required to flesh out a complete memory system? We do think a HAL-like model that was sensitive to the same co-occurrences in the natural environment as a human-language learner (not just the language stream) would be able to capitalise on this additional information and construct more meaningful representations.

Subconceptual Representations. Since HAL's vector representations seem to carry such a board range of information (semantic, grammatical and thematic), we think that the representations are best considered subconceptual—a set of distributed units that can represent much of what one would want a mental representation to do. Much has been made of the advantages of subconceptual representations (Cussins, 1990; Smolensky, 1988; Van Orden, Pennington, & Stone, 1990). For example, the concept *bachelor* could be represented by having the features *adult*, *single* and *male*

activated. However, our view is that subconceptual representation is not just some set of semantic features because such a conception begs two questions. First, how does one justify these particular features as subconceptual? Second, one can argue that microfeatures such as these are just more thinly sliced symbolic features (to paraphrase Cussins's characterisation of Fodor & Pylyshyn, 1988). The features used by HAL (the vector elements) are discrete aspects of the input stream that when encoded form a distributed representation that behaves in a variety of cognitively meaningful ways. No commitments are required to a particular set of features, and the encoding scheme is tightly tied to the actual input. This, together with how the vector representations are able to account for such a broad range of cognitive phenomena, makes this type of memory vector a good candidate for a subconceptual representation.

CONCLUSIONS

An increasing role for contextual information during comprehension is consistent with the general constraint satisfaction approach (MacDonald et al., 1994). We suspect that global co-occurrence models, more so than local co-occurrence models, will better capture the richness of cognitive and language effects that are important to the comprehension process. Landauer and Dumais (in press) have shown that their LSA model can mimic synonym vocabulary-test performance and also account for semantic context effects in lexical ambiguity. Earlier work with HAL has demonstrated that the representations can account for a broad range of cognitive effects, including semantic priming (Lund & Burgess, 1996a; Lund et al., 1995, 1996), basic categorisation (Lund & Burgess, 1994), cerebral asymmetries (Burgess & Lund, 1996) and the effects of brain damage on the semantic system (Buchanan et al., 1996). In this paper, we argue that HAL's contextual representations can provide a rich computational measure of semantic and grammatical effects.

REFERENCES

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain and Cognition*, *32*, 111–114.

Burgess, C. (1991). *Interaction of semantic, syntactic and visual factors in syntactic ambiguity resolution.* Unpublished doctoral dissertation, University of Rochester, Rochester, NY.

Burgess, C., & Hollbach, S.C. (1988). A computational model of syntactic ambiguity as a lexical process. In *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society*, pp. 263–269. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Burgess, C., & Lund, K. (1994). Multiple constraints in syntactic ambiguity resolution: A connectionist account of pyscholinguistic data. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, pp. 90–95. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Burgess, C., & Lund, K. (in press). Modeling cerebral asymmetries of semantic memory using high-dimensional semantic space. In M. Beeman & C. Chiarello (Eds), *Getting it right: The cognitive neuroscience of right hemisphere language comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Burgess, C., & Tanenhaus, M.K. (1996). *Semantic, syntactic and visual factors in syntactic ambiguity resolution.* Unpublished manuscript, University of California Riverside.

Burgess, C., Livesay, K., & Lund, K. (1996). Modeling parsing constraints in high-dimensional semantic space: On the use of proper names. In *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society*, p. 737. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Clifton, C., & Ferreira, F. (1989). Ambiguity in context. *Language and Cognitive Processes*, *4*, 77–103.

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

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Cummins, R. (1996). *Representations, targets, and attitudes.* Cambridge, MA: MIT Press.

Cussins, A. (1990). The connectionist construction of concepts. In M.A. Boden (Ed.), *The philosophy of artificial intelligence*, pp. 367–440. Oxford: Oxford University Press.

Deese, J. (1965). *The structure of associations in language and thought*, pp. 97–119. Baltimore, MD: Johns Hopkins Press.

Dyer, M.G. (1990). Distributed symbol formation and processing in connectionist networks. *Journal of Experimental and Theoretical Artificial Intelligence*, *2*, 215–239.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Ervin, S.M. (1961). Changes with age in the verbal determinants of word-association. *American Journal of Psychology*, *74*, 361–372.

Ervin, S.M. (1963). Correlates of associative frequency. *Journal of Verbal Learning and Verbal Behavior*, *1*, 422–431.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories by unsupervised learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, pp. 820–825. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory and Cognition*, *5*, 335–339.

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

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* PhD thesis, University of Connecticut, Storrs, CT. Distributed by Indiana University Linguistics Club.

Frazier, L., & Clifton, C. (1996). *Construal.* Cambridge, MA: MIT Press.

Frazier, L., & Fodor, J.D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.

Gallant, S.I. (1991). A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, *3*, 293–309.

Gernsbacher, M.A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Glenberg, A. (in press). What memory is for. *Behavioral and Brain Sciences.*

Gorrell, P. (1989). Establishing the loci of serial and parallel effects in syntactic processing. *Journal of Psycholinguistic Research*, *18*, 61–74.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics* , *19*, 103–120.

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

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

Irmscher, W.F. (1972). *The Holt guide to English*. New York: Holt, Rinehart and Winston.

Juliano, C., & Tanenhaus, M.K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, pp. 593–598. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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.

Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500–526.

Kurtzman, H.S. (1985). On-line probe recognition latencies within complex sentences. *Language and Speech*, *28*, 143–156.

Landauer, T.K., & Dumais, S. (1994). Memory model reads encyclopedia, passes vocabulary test. Paper presented at the *Psychonomics Society Meeting*, St. Louis, MO, November.

Landauer, T.K., & Dumais, S.T. (in press). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*.

Lund, K., & Burgess, C. (1994). Extraction of semantics via lexical co-occurrence: Methodology and preliminary results. Paper presented at the *NIPS*94 (Neural Information Processing Systems) Statistical and Neural Network Approaches to Natural Language Processing Post-conference Workshop*, Vail, CO, December.

Lund, K., & Burgess, C. (1996a). Hyperspace Analogue to Language (HAL): A general model of semantic representation (abstract). *Brain and Cognition*, *30*, 265.

Lund, K., & Burgess, C. (1996b). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203–208.

Lund, K., Burgess, C., & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society*, pp. 660–665. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Lund, K., Burgess, C., & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society*, pp. 603–608. Hillside, NJ: Lawrence Erlbaum Associates Inc.

Lupker, S.J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, *23*, 709–733.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*, 157–201.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Mandler, G. (1962). From association to structure. *Psychological Review*, *69*, 415–427.

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

McClelland, J.L., & Kawamoto, A.H. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*, pp. 272–325. Cambridge, MA: MIT Press.

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

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, *20*, 155–171.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–235.

Moss, H.E., Hare, M.L., Day, P., & Tyler, L.K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*, 413–427.

Moss, H.E., Ostrin, R.K., Tyler, L.K., & Marslen-Wilson, W.D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 863–883.

Neely, J.H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G.W. Humphreys (Eds), *Basic processes in reading: Visual word recognition*, pp. 264–336. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin*, *84*, 93–116.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Pearlmutter, N.J., & MacDonald, M.C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language*, *34*, 521–542.

Plaut, D.C., & Shallice, T. (1994). *Connectionist modelling in cognitive neuropsychology: A case study*. Hove: Lawrence Erlbaum Associates Ltd.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 358–374.

Saint-Dizier, P., & Viegas, E. (Eds) (1995). *Computational lexical semantics*. Cambridge: Cambridge University Press.

Schutze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92*, pp. 787–796. Los Alamitos, CA: IEEE Computer Society Press.

Schvaneveldt, R.W. (Ed.) (1990). *Pathfinder associative networks: Studies in knowledge organizations*. Norwood, NJ: Ablex.

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

Shelton, J.R., & Martin, R.C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 1191–1210.

Slator, B.M., Amirsoleymani, C., Andersen, S., Braaten, K., Davis, J., Ficek, R., Hakimzadeh, H., McCann, L., Rajkumar, J., Thangiah, S., & Thureen, D. (1990). Towards empirically derived semantic classes. In *Proceedings of the Fifth Annual Rocky Mountain Conference on Artificial Intelligence*, pp. 257–262. Las Cruces, NM.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*, 143–177.

Smith, E.E., Shoben, E.J., & Rips, L.J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241.

Smith, F. (1971). *Understanding reading*. New York: Holt.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–74.

Spence, D.P., & Owens, K.C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317–330.

St. John, M.F., & McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

Tanenhaus, M.K., & Carlson, G.N. (1989). Lexical structure and language comprehension. In W. Marslen-Wilson (Ed.), *Lexical representation and process*, pp. 529–561. Cambridge, MA: MIT Press.

Toglia, M.P., & Batig, W.F. (1978). *Handbook of semantic word norms.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Trueswell, J.C., Tanenhaus, M.K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 528–553.

Trueswell, J.C., Tanenhaus, M.K., & Garnsey, S.M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318.

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