Behavior Research Methods, Instruments, & Computers
1998, 30 (2), 188-198

INVITED ADDRESS

From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model

CURT BURGESS
University of California, Riverside, California

This paper presents a theoretical approach of how simple, episodic associations are transduced into semantic and grammatical categorical knowledge. The approach is implemented in the hyperspace analogue to language (HAL) model of memory, which uses a simple global co-occurrence learning algorithm to encode the context in which words occur. This encoding is the basis for the formation of meaning representations in a high-dimensional context space. Results are presented, and the argument is made that this simple process can ultimately provide the language-comprehension system with semantic and grammatical information required in the comprehension process.

> ... "meaning" seems to connote, for most psychologists at least, something inherently nonmaterial, more akin to "idea" and "soul" than to observable stimulus and response, and therefore to be treated like the other "ghosts" that J. B. Watson dispelled from psychology." (Osgood, Suci, & Tannenbaum, 1957, p. 1)

There are many ways to characterize meaning (Ogden & Richards, 1923). Lexical semantics usually refers to the meanings of individual words and distinctions constrained by their morphology. Structural semantics involves the integration of these word meanings into a more complex meaning guided by the syntax of the sentence. Of course, we do not tend to remember sentences; we abstract this level of meaning to what can be referred to as a *mental model* of a situation in which this linguistic information—lexical-semantic, structural, and/or our experience of the environment—combine to form a complex mental representation that is not limited to the linguistic form (see Garnham, 1996; Johnson-Laird, 1983). The purpose of this paper is to describe a new class of memory models that are referred to as *high-dimensional* memory models.[1] These models have as their goal extracting and representing meaning from a stream of language. The nature of the meaning that can be extracted with these models is at the level of the word and is to some extent, relevant to larger units of meaning. Although computational high-dimensional memory models can be considered "new," the fact is that they have many connections to past and present modeling efforts. Charles Osgood (Osgood, 1971; Osgood et al., 1957) should be considered the father of high-dimensional memory models. These models also have at their heart a clear associationist bent (Deese, 1965). Another more recent functional precursor to high-dimensional memory models is Elman's (1990) work with recurrent neural networks that learn generalized representations by virtue of their experience with input from the environment. Elman's work has operationally defined the message about how context is important.

At the same time, however, there are limitations to these historical precursors that set the stage for the advantages of high-dimensional memory models. Osgood et al.'s (1957) semantic differential approach required vast numbers of human judgments about the nature of words in order to derive a set of semantic dimensions. In addition, the semantic differential technique (although very useful) requires one to commit to a set of semantic features upon which all words will literally be judged. This is a very tricky enterprise, which makes very strong presumptions about how the language user interacts with the environment and encodes experience. Deese's (1965) work hinged on the use of word association norms that correspond, in some sense, to meaning to provide a set of intercorrelations among words. Both word-association norms and the semantic differential technique require considerable human overhead to gather word-meaning information for even a small set of items. It is also the case that neither of these approaches offers any model of how information gets organized in the first place. In contrast, connectionist learning models have a completely different type of limitation. It would simply take an unreasonable amount

This research was supported by NSF Presidential Faculty Fellow Award SBR-9453406 to the author. A version of this paper was presented as the keynote talk at the 1997 Society for Computers in Psychology (SCiP) annual meeting in Philadelphia. Kevin Lund, Catherine Decker, and two anonymous reviewers provided many helpful comments. More information about research at the Computational Cognition Lab, a HAL demo, and reprint information can be found at http://HAL.ucr.edu. Correspondence should be addressed to C. Burgess, Psychology Department, 1419 Life Sciences Bldg., University of California, Riverside, CA 92521-0426 (e-mail: curt@cassandra.ucr.edu).

of computer-processing time to develop a model of memory from the combination of a very large corpus and a set of lexical items used to support human language. The time results from the complexity of the learning algorithms and the number of passes through a corpus that a connectionist model requires to complete its learning. Fortunately, at a functional level, we have demonstrated that recurrent neural networks and our global co-occurrence learning procedures are both learning models that utilize a word's context to induce word meaning and, given the same input, will produce very similar results (Lund & Burgess, 1997).

This paper describes a particular high-dimensional memory model, the hyperspace analogue to language (HAL) model. The model makes completely explicit the nature of associations and the relationship of associations to categorical knowledge. As a result, this model can address a variety of difficult and problematic theoretical issues in learning and memory. This will occur in the context of describing how global lexical co-occurrence (the learning procedure that HAL uses) can provide some of the basic building blocks of a language-comprehension system.

THE HAL MEMORY MODEL

Having a plausible methodology for representing word meaning is an important component to a model of memory, particularly when the goal is to extend the model to subserve language processing. HAL uses a large corpus of text, and the basic methodology involves tracking lexical co-occurrences within a 10-word moving window that "slides" along the text. Using these co-occurrences, a high-dimensional meaning space of 140,000 lexical dimensions can be developed. Since each vector element represents a symbol (usually a word) in the text input, we refer to this high-dimensional meaning space as a *context* space. This high-dimensional context (or meaning) space is the memory matrix that can be used to simulate experiments or to further analyze word meanings.

Constructing the Memory Matrix

As mentioned before, the basic methodology of the model involves developing a matrix of word co-occurrence values for a set of words by moving a 10-word window along the corpus of text. Within the moving window, weighted lexical co-occurrence counts are tabulated and are inversely proportional to the number of words separating a pair of words. Words that are closer together in the moving window get a larger weight. Table 1 shows an example of the matrix-construction procedure for a five-word window using the example sentence "the horse raced past the barn fell." The matrix is actually two triangular matrices folded into one table—that is, rows encode co-occurrences in the window that occur before a word, and columns encode co-occurrence information occurring after the word. Consider, for example, the word *barn* in this sentence. Co-occurrence values of words occurring prior to *barn* are in the *barn* row. The word "past" is separated by one word from *barn* and thus gets a 4. If *past* had occurred adjacent to *barn*, it would have received a 5. There are two occurrences of the word *the* in the example sentence, and there is a 6 recorded for *barn* (row) and *the*. One occurrence of *the* occurs just prior to *barn* and gets a 5. The other *the* in this sentence is five words away and get a 1. The 6 in this cell represents the addition of the five and one co-occurrence values. Columns work the same way (to encode the subsequent co-occurrences), except that the point of reference is from the column word. This example was made using a five-word window; however, a 10-word window was used in the actual model, and HAL's matrix is 70,000 square items.

The corpus that served as input to the HAL model is approximately 300 million words of English text gathered from Usenet newsgroups that contained English text. Properties of Usenet text that were appealing were both its conversational nature and its diverse nature, making it closer in form to everyday speech. An important goal in this project is to develop a model that does minimal preprocessing of the input. It is to HAL's credit that it works with this noisy, conversational input, thus managing some of the same problems that the human-language comprehender encounters.

Matrix Properties

HAL's vocabulary consists of the 70,000 most frequently used symbols in the corpus. About half of these symbols had entries in the standard Unix dictionary. The remaining items were nonword symbols, misspellings, proper names, and slang. These atypical items were kept in the lexicon since later research or applications may require the use of vectors of misspellings, proper names, and so on. The co-occurrence procedure produces a 70,000 × 70,000 matrix similar to (and much larger than) the one shown in Table 1. Recall that each row of the matrix represents the degree to which each word in the vocabulary precedes the word corresponding to the row. Similarly, each column represents the co-occurrence values for words following the word corresponding to the column.

Table 1
Sample Global Co-occurrence Matrix for the Sentence "the horse raced 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—The values in the matrix rows represent co-occurrence values for words that preceded the word (row label). The values in the columns represent co-occurrence values for words following the word (column label). Cells containing zeros were left empty in this table. This example uses a five-word co-occurrence window.

Vector Properties

A full co-occurrence vector for a word consists of a concatenation of the row and the column for that word. Thus, the vector for *barn* (from Table 1) would be 0, 2, 4, 3, 6, 0, 5, 0, 0, 0, 0. The vector for *barn* in the full matrix would be 140,000 elements long rather than 10 elements long as in this example.

We view these vectors as coordinates of points in a high-dimensional space, with each word occupying one point. Each vector corresponds to a lexical symbol in the input stream. These symbols are not always words (they may be numbers, emoticons, etc.), although it is convenient to refer to them as word vectors. In most simulations, vectors of 140,000 elements in length are used; however, one should not conclude that we think it best to characterize human memory as represented by 140,000 dimensions. We suspect that much less dimensionality is required. For our work, only 100–200 vector elements seem necessary. Some effects can be easily carried by as few as 10 vector elements, which is useful in connectionist models where the training time is often nonlinearly related to vector size. What is important in any attempt to shorten a vector is to keep the vector elements that are most "informative." To determine this, the column and row variance is calculated, and the elements with the smallest variance are not used. Variance as an information measure is useful in a model like HAL since it is a direct reflection of the diversity of the contexts in which words are found. Approximately 100–200 vector elements contain most of the variance in these word vectors. For example, a 200-element word vector would not use the lowest variant 139,800 vector elements. In most of our simulations, however, we use the full vector simply because it is computationally more straightforward than computing variance across such a large matrix of numbers.

Using this type of vector representation, differences between two words' co-occurrence vectors (e.g., word x_i and word y_i) can be measured as the distance between the high-dimensional points defined by their vectors. Distance between two words can be computed using a similarity metric. We typically use a Minkowski metric (see Equation 1) and usually use a Euclidean distance metric ($r = 2$).

$$\text{distance} = \sum \langle | x_i - y_i |^r \rangle^{1/r}. \quad (1)$$

Each element of a vector represents a coordinate in high-dimensional space for a word or concept, and a distance metric applied to these vectors presumably corresponds to context (not just item) similarity. The vectors can also be viewed graphically as can be seen in Figure 1. Sample words (e.g., *france*, *puppy*) are shown with their accompanying 25-element vectors (the 25 most variant of the 140,000 elements are shown for viewing ease). Each vector element has a continuous numeric value (the frequency normalized value from its matrix cell).[2] A gray scale is used to represent the normalized value, with white corresponding to a zero or minimal value. The word vectors are very sparse; a large proportion of a word's vector elements are zero or close to zero.

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 representation 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 1,000 elements. Finally, it can be seen that words representing similar concepts have similar vectors, although this can be subtle at times (Figure 1). See Lund and Burgess (1996) 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 (symbols in the input stream). This translates into the ability to have vector representations for abstract concepts (e.g., *justice*, *reality*) as easily as one can have representations for more basic concepts (e.g., *dog*, *book*) (Burgess & Lund, 1997b). This is important, if not absolutely crucial, when developing a memory model that purports to be general in nature and that provides the essential bottom-up input to a language-comprehension system.

CONSTRUCTING THE BUILDING BLOCKS OF LANGUAGE

The goal of language understanding is to derive meaning from some sequence of words. In building a language-comprehension model, one has to describe how word-level categorical information is formed. The HAL model encodes categorical information, both semantic and grammatical. These two sources of information provide the front end to the ability of combining words into larger meaningful units.

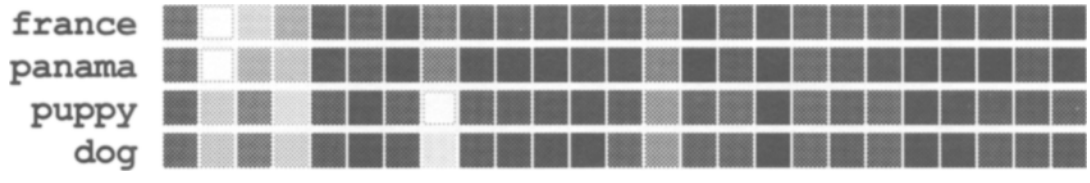

Figure 1. Sample 25-element word vectors for four words. Each vector element has a continuous value (the normalized value from its matrix cell) and is gray-scaled to represent the normalized value with white corresponding to zero.

Semantic Knowledge

There are numerous models of semantic memory, and, as one might expect, these models entail considerable disagreement over countless details. However, there are several commonly held assumptions. Semantic models tend to focus on words and their relationship to other words, sense rather than reference. These relationships also tend to be summed across experience, rather than representing specific episodes. At the same time, these semantic memories are constructed as a function of our episodic experience with the world, and they allow for the categorization of information.

Similarly, there are many ways in which cognitive psychologists can assess the structure and processing components of semantic memory. There are three ways in which we have investigated semantic structure with the HAL model, often in conjunction with experiments with human subjects. First, HAL's word vectors seem to possess sufficient information from the global co-occurrence procedure to support basic semantic categorization. Second, we have compared distances in the high-dimensional memory space with the priming data that can be obtained by using the semantic priming methodology. This work has implications for how we think about word associations. Lastly, the area in the high-dimensional space surrounding a word contains other words that constitute a context (or a semantic) neighborhood for that word. We will review evidence that suggests that HAL's context neighborhoods possess certain denotative characteristics of word meaning.

Semantic categorization. Using human judgments about concepts to determine categorical structure has a long history in cognitive psychology. In their well-known study, Rips, Shoben, and Smith (1973) used typicality ratings of different kinds of birds to generate a set of predictions for the semantic verification model. They then used a multidimensional scaling (MDS) procedure to transform these ratings into a two-dimensional pictorial representation of these different types of birds, which demonstrated that the typicality ratings provided important information for categorization purposes. With HAL, rather than human ratings, the word vectors can be used to categorize concepts. The vectors represent the contextual history of a word, and we have seen that the distances between these vectors are meaningful from the standpoint of word neighborhoods and semantic priming. In Figure 2, we see that categorical structure emerges when we view the MDS solution for body parts, animals, and geographic locations. These three groups of words segregate nicely.[3] It is important to note that all the MDS presentations presented in this paper are a reduction of 140,000 dimensions to two dimensions, which results in a loss of resolution in the categorical integrity that can be conveyed. Still, it is clear that the word vectors allow for categorization much like human ratings of similarity. The objects in this MDS are all concrete nouns. Other categorization work has shown that HAL's representations extend to the categorization of abstract nouns and emotional words (Burgess & Lund, 1997b).

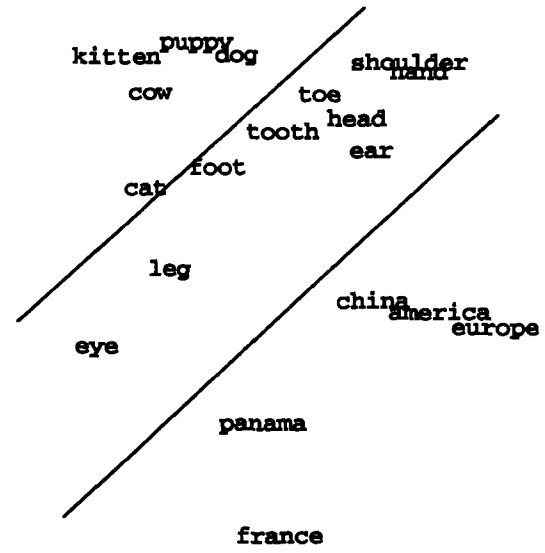

Figure 2. Two-dimensional multidimensional scaling solution for animals, body parts, and geographic locations.

Semantic neighborhoods. The meaning space in HAL is a high-dimensional space in which each word is represented as a point in space as a function of the vector coordinates. The distances that these words are from one another are results of the contexts in which they systematically occurred. Similar words occur in similar contexts. As a result, one should be able to see semantic and contextual relationships among the words that are close to another word in the hyperspace. We have found that these neighborhoods are rich in meaningfully related concepts. Table 2 shows the six closest neighbors for two words: *beatles* and *frightened*. Inspection suggests that the notion of a musical group is inherent in the neighborhood for *beatles*. Other neighbors of *beatles* further constrain the concept (*best, british, greatest*). The nature of these neighborhoods is more connotative than denotative. Some of *beatles* neighbors are concepts that are definitional (i.e., *beatles* is a *band*, is likely considered by many as *original*, and has produced many *songs* and *albums* [and *movies*]). The semantic neighborhood is more of a definition by implication. It certainly is not a typical denotative definition such as "English quartet of composers and musicians; members are ..." (Morris, 1971). The neighbors represent a set of constraints on meaning. Some neighbors may be almost synonymous; other neighbors characterize various aspects of meaning. Meaningful neighborhoods are not limited to nouns. Table 2 also contains the neighborhood for *frightened* and is similarly informative providing synonyms and words related to other emotional aspects of *frightened*. These neighbors are the closest set of words to *frightened*. The nature of the meaning acquisition process sug-

Table 2
Example Neighborhoods for *beatles* and *frightened*

Word	Neighbors
beatles	*original*
	band
	song
	movie
	album
	songs
frightened	*scared*
	upset
	shy
	embarrassed
	anxious
	worried

gests that these neighbors are used in contexts similar to the contexts in which *frightened* is found. Humans are able to generate a more precise and more elaborate denotative definition for a word than these context neighborhoods. HAL likely provides a good approximation of the initial activation of meaning; denotative definitions certainly require considerable top-down processing—a component that this representational model does not have.

At the same time, however, these context neighborhoods seem to provide a sufficient set of cues such that humans can generate the source word or a word very close in meaning to it. We have systematically investigated these types of neighborhoods (Burgess, Livesay, & Lund, 1998; Lund & Burgess, 1996). One approach we have taken in evaluating the cognitive plausibility of these neighborhoods is to present people with the neighborhoods and see whether or not they can determine what the target concept is. We found that about 20% of the time people generate the specific word that we used to extract the neighbors. Even when we eliminated these "direct" hits from the data analysis, the remaining trials show that subjects are able to reliably use the neighborhoods to produce something semantically close to the target. The neighborhoods suggest that retrieval in a model such as HAL could produce a set of related items that could be useful to higher level cognitive systems during language comprehension (Burgess et al., 1998).

Semantic priming. The meaningful nature of the neighborhoods suggests that distances in the high-dimensional space would correspond to meaning activation in other cognitive tasks. Priming is a widely used approach to investigate memory organization and meaning retrieval. The most common technique used in investigating semantic priming is the single-word priming procedure (see Neely, 1991, for a review). Semantic priming occurs when a prime, such as *dog*, facilitates a target, such as *cat*, presumably due to activation spreading in memory from the prime to the target. Responses (lexical decisions or pronunciation) to the primed target are quicker than responses to a target preceded by an unrelated word.

There has been an ongoing controversy in the priming literature as to what is meant by "semantic" priming and under what conditions it is obtained. Critical to this discussion is a distinction between semantic and associative relationships. In most experiments, word-association norms are used to derive stimuli. However, word norms confound semantic and associative relationships. *Cat* and *dog*, in the example above, are related both categorically (are similar animals) and associatively (one will tend to produce the other in production norms). The typical assumption behind associative relationships is that associations are caused by temporal co-occurrence in language (or elsewhere in the environment). Stimuli can be constructed such that these semantic-categorical and associative relationships can be, for the most part, orthoginally manipulated. To illustrate, cat and dog are semantically and associatively related. However, *music* and *art* are semantically related, but *art* does not show up as an associate to *music* in word norms. Conversely, *bread* tends to be one of the first words produced in norms to the word *mold*. However, *bread* and *mold* are not similar. Clearly, though, this is not to say there is no relationship between *bread* and *mold*; they are just very different items. As the story goes, *mold* and *bread* would be likely to co-occur. Examples of these types of word pairs can be seen in Table 3.

Our claim is that HAL encodes experience such that it learns concepts more categorically. Associative (more episodic) relationships will have been aggregated into the conceptual representation. This can be seen by examining Table 1. The vector representation for *barn* will include the row and column of weighted co-occurrence values for the words that co-occurred with *barn* in the moving window. The representation for *barn*, as it stands in Table 1, is episodic. *Barn* has occurred in only this one context. As more language is experienced by HAL, the vector representation for *barn* accrues more contextual experience; as a result, the weighted co-occurrences sum this experience, resulting in a more generalized representation for *barn*. This is an important aspect of HAL for attempting to model priming. It follows that the distances in the hyperspace should be sensitive to more generalized, categorical relationships. Furthermore, the more associative relationships should not have a strong correlation to HAL's distance metric. We tested these hypotheses in two experiments (Lund, Burgess, & Atchley,

Table 3
Example Prime–Target Word Pairs From the Semantic, Associated, and Semantic + Associated Relatedness Conditions

Condition	Word Pairs
Semantic	*table–bed*
	music–art
	flea–ant
Associated	*cradle–baby*
	mug–beer
	mold–bread
Semantic + Associated	*ale–beer*
	uncle–aunt
	ball–bat

Note—The full set of these stimuli was taken from Chiarello, Burgess, Richards, and Pollock (1990).

1995) using the three different types of word relationships illustrated in Table 3. These word relationships have various combinations of semantic and associative properties: semantic only, associative only, and combined semantic and associative properties. There is considerable research that shows that human subjects are sensitive to all three of these types of word relationships (Lund et al., 1995; Lund, Burgess, & Audet, 1996; see Neely, 1991). We replicated that finding: Subjects made faster lexical decisions to related word trials (in all three conditions) than to the targets in the unrelated pairs (Lund et al., 1995). In a second experiment, we computed the semantic distance between the related and unrelated trials in all three conditions using HAL. Priming would be computed in this experiment by using the distances; there should be shorter distances for the related pairs than for the unrelated pairs in the representational model. In this experiment, we found robust priming for the semantic-only and the semantic-plus-associative conditions. There was no distance priming in the model for the associated-only pairs. This result raises some intriguing questions about the representational nature of words and the ongoing controversy in the priming literature as to what is meant by "semantic" priming and under what conditions it is obtained.

The controversy exists, in part, due to a mixed set of results in the literature: some investigators obtaining semantic priming without association, others not finding semantic-only priming in conditions that would seem to limit strategic processing. Fischler (1977) has one of the earliest findings showing that strength of association did not correlate with priming. Similarly, Chiarello, Burgess, Richards, and Pollock (1990) found semantic-only priming using a low proportion of related trials and a naming task. However, Lupker (1984) did not find priming for semantically related word pairs that were not also associatively related. A similar set of results is found in Shelton and Martin (1992). They used a single presentation lexical decision task where words were presented one after another with lexical decisions made to each word. Such a procedure masks the obviousness of prime–target relations to a subject. Shelton and Martin did not find semantic priming under these conditions. A comparison of experiments such as these usually entails a comparison of the methodologies. Experiments that do not obtain semantic-only priming typically avoid the lexical decision task, unless it is part of the individual presentation procedure (i.e., Shelton & Martin, 1992). The naming task is thought to be less sensitive to strategic effects (although this may also limit its sensitivity to semantic relations). Clearly, experimental procedures and task differences play a part in these results. Focusing on task differences, however, may divert attention from important representational issues that are likely to be just as important. In developing representational theory, it is important not to make representational conclusions based solely on procedural issues.

We have argued that an experiment's sensitivity in reflecting the semantic-only priming effect is guided by the strength of the semantic (contextual) relationship. One set of stimuli that we have evaluated in detail using the HAL model is that used by Shelton and Martin (1992). We found that many of their semantic pairs (e.g., *maid–wife*, *peas–grapes*) were not closely related by using HAL's semantic distance metric. Furthermore, a number of their semantic and associated pairs were very strongly related categorically (e.g., *road–street*, *girl–boy*) (see Lund et al., 1995). Using HAL, we argued that the semantic-only condition did not produce priming simply because the prime–target pairs in that condition were not sufficiently similar.

There are two experiments that offer compelling evidence that increased similarity results in priming under task constraints usually associated with a lack of semantic-only priming. Cushman, Burgess, and Maxfield (1993) found priming with the semantic-only word pairs used originally by Chiarello et al. (1990) with patients who had visual neglect as a result of brain damage. What is compelling about this result is that the priming occurred when primes were presented to the impaired visual field. These patients were not aware that a prime had even been presented, thus making it difficult to argue for any strategic effect. A more recent result by McRae and Boisvert (in press) confirmed our earlier hypothesis generated by our HAL simulation that Shelton and Martin's (1992) failure to find priming was due to insufficient relatedness in their semantic-only condition. Recall that they used an individual-presentation lexical decision methodology. McRae and Boisvert replicated this methodology but used a set of nonassociatively related word pairs that subjects rated as more similar than Shelton and Martin's items. McRae and Boisvert replicated Shelton and Martin with their items but, using the more similar items, found a robust semantic-only priming effect. Thus, it appears that increased attention to the representational nature of the stimuli affords a more complete understanding of the semantic constraints as well as the methodological issues involved in priming.

HAL's distance metric offers a way to evaluate stimuli in a clearly operationalized manner. The large lexicon provides the basis for which the stimuli from various experiments can be evaluated directly. In most experiments, word association norms are used to derive stimuli, and it is important to realize that word norms confound semantic and associative relationships.

We argue that HAL offers a good account of the initial bottom-up activation of categorical information in memory. It provides a good index of what information can be activated automatically. Although others (Lupker, 1984; Shelton & Martin, 1992) have argued that it is associative, not semantic, information that facilitates the automatic, bottom-up activation of information, some of the confusion is a result of the field not having a clear idea of what an association is or represents. On one hand, an association is operationally defined as the type of word re-

lationships that is produced when a person free associates. Yet this is an unsatisfying definition at a theoretical level. It also confounds many types of word relationships that can be found using a word-association procedure. One intuitive conception of word association is that it is related to the degree to which words tend to co-occur in language (Miller, 1969). Spence and Owens (1990) confirmed this long-held belief empirically. To see whether this relationship between word association ranking and lexical co-occurrence held for the language corpus that we use for HAL, we used 389 highly associated pairs from the Palermo and Jenkins (1964) norms as the basis for this experiment (Lund et al., 1996). We replicated Spence and Owens's effect; word association ranking was correlated (+.25) with frequency of co-occurrence (in the moving window). Our correlation was not as strong as theirs, probably due to the fact that we used only the five strongest associates to the cue word. However, using all strongly associated word pairs allowed us to test a further question. To what extent is similarity, at least as operationalized in the HAL model, related to this co-occurrence in language for these highly associated words? We divided these strongly associated pairs into those that were semantic neighbors (associates that occurred within a radius of 50 words in the hyperspace) and those that were nonneighbors (pairs that were farther than 50 words apart). Since all these items are strong associates, one might expect that the word-association ranking should correlate with co-occurrence frequency for both HAL's neighbors and HAL's nonneighbors (recall that these two groups of words collectively show a +.25 correlation between ranking and co-occurrence). The results were striking. The correlation using the close neighbors is +.48; the correlation for the nonneighbors is +.05. These results suggest that the popular view that association is reflected by word co-occurrence seems to be true *only* for items that are similar in the first place. Word association does not seem to be best represented by any simple notion of temporal contiguity (local co-occurrence). From the perspective of the HAL model, word meaning is best characterized by a concatenation of these local co-occurrences (i.e., global co-occurrence—the range of co-occurrences [or the word's history of co-occurrence]) found in the word vector. A simple co-occurrence is probably a better indicator of an episodic relationship, but a poor indicator for more categorical or semantic knowledge. One way to think about global co-occurrence is that it is the contextual history of the word. The weighted co-occurrences are summed indices of the contexts in which a word occurred. Next, we will review evidence that the contextual nature of the origin of meaning in HAL's representations provides more than simple semantic knowledge.

Grammatical Knowledge

In the previous section, the high-dimensional meaning spaces were described as both semantic and contextual. The bulk of the early work with the HAL model investigated issues that were typically semantic in nature: various types of semantic and associative priming, use of neighborhoods as word definitions, semantic paralexias, and accounting for results of word production norms. The simple idea behind the HAL model (and certainly not unique to our work) that context determines word meaning has been presented in another linguistic domain. Ervin (1963; Ervin-Tripp, 1970) and many others (see Nelson, 1977) have shown that a child's experience with context is implicated in the acquisition of both semantic and grammatical knowledge.

There is compelling evidence that the word vectors encoded in HAL's global co-occurrence procedure carry information about grammatical class. Perhaps it is not surprising that the contextual nature of the representations could encode grammatical information, given the importance of context in a developing child's grammatical competence. At the same time, however, a singular acquisition mechanism that can learn semantic as well as grammatical-class information violates the traditionally held notion of representational modularity (see Burgess & Lund, 1997a). We have explored these grammatical and syntactic representational issues in detail elsewhere (Burgess et al., 1998; Burgess & Lund, 1997a; see Finch & Chater, 1992, for a similar approach).

In this section, two analyses of grammatical class will be presented that show that HAL's vector representations encode grammatical-class information as well as semantic information. Vectors of words of different grammatical classes were extracted from the model and analyzed using an MDS procedure, just as with the semantic categorization presented above. In Figure 3, it can be seen that the nouns, verbs, determiners, and prepositions cluster into their own spaces. Words that are noun–verb am-

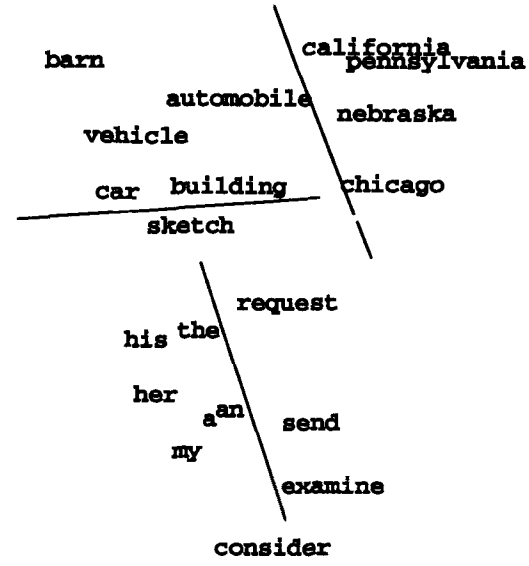

Figure 3. Two-dimensional multidimensional scaling solution for nouns, locations, verbs, and determiners.

biguities (*sketch, building*) straddle the high-dimensional fence between these two grammatical categories. The same pattern of grammatical-class discrimination was seen in Burgess and Lund (1997a), in which a more systematic approach was taken. Nouns and verbs that had been part of the stimuli from an online parsing experiment were used, along with a larger set of determiners and prepositions (Simulation 2). The results presented here replicate the basic grammatical categorization effect found in the Burgess and Lund (1997a) paper.

Nouns, verbs, prepositions, and determiners occur in very different parts of a sentence due to their differing roles in the syntactic structure of the sentence. From the perspective of contextual constraints and how substitutability will play a role in these contextual constraints, it follows that these coarsely coded grammatical categories would be part of what is encoded in a global co-occurrence procedure such as HAL's. A much more challenging problem would be the categorization of subcategories of classes all belonging to the same grammatical category. Figure 4 illustrates this with different types of determiners. Articles (*a, an, the*), demonstratives (*this, these, those*), and genitives (*her, his, our, their*) all separate into their own portion of the context space. This result goes beyond the earlier result that demonstrated the separation of the distinct grammatical classes. In this case, these determiners occupy the same location in the surface structure of a sentence: *a dog, this dog, her dog*. However, the higher order distributional information associated with these different classes of determiners has to be implicit in the word vectors in order to obtain a result such as that shown in Figure 4. A similar pattern can be seen when we look at the high-dimensional spaces occupied by simple past-participle verbs and past-tense verbs, such as those seen in Figures 5A and 5B, respectively (modeled after

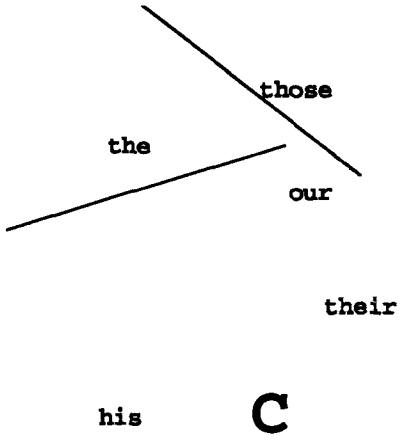

Figure 4. Two-dimensional multidimensional scaling solution for three subclasses of determiners: (A) articles, (B) demonstratives, and (C) genitives.

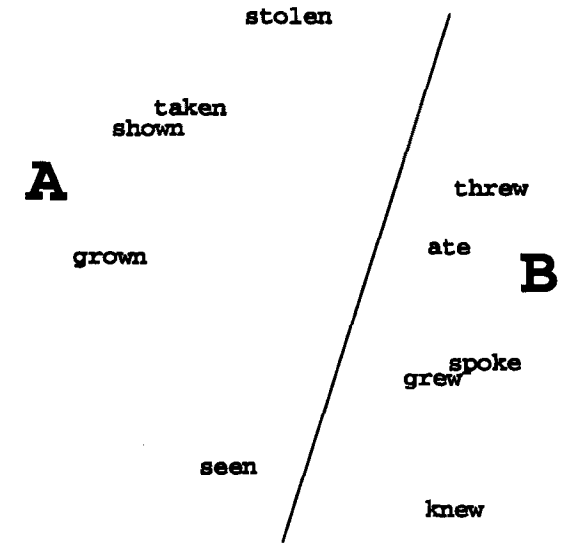

Figure 5. Two-dimensional multidimensional scaling solution for (A) unambiguous past-participle verbs and (B) unambiguous past-tense verbs.

Burgess & Lund, 1997a, Simulation 3). These verbs are unambiguous and clearly distinguished by their vector representations. The exception to this is the past participle verb *seen*. Although *seen* is a past participle, it is coming into relatively common, albeit ungrammatical, use as a past tense (as in "I seen it"), which is reflected in the input corpus and, ultimately, in the word's vector representation and its migration to the past-tense space. In Burgess et al. (1998), an analogous pattern of results was seen with verbs of implicit causality, such as *flatter* and *praise*. Verbs of implicit causality are interpersonal verbs that have a semantic bias that specifies whether the agent or the patient is most likely to carry out the action of the verb. This becomes clear in Sentences 1a and 1b.

1a. The sportscaster praised the Huskers.

1b. The sportscaster flattered the Wolverines.

In Sentence 1a, the sportscaster is praising a team because of their impressive victory in the Orange Bowl. In Sentence 1b, the verb *flattered* suggests an ulterior motive on the part of the sportscaster—probably a disingenuous attempt to interview players after a narrow victory in the Rose Bowl. Au (1986) had subjects sort verbs of implicit causality into "meaningful" groups. The MDS of the human categorization reflecting Au's results and the MDS of these verbs in Burgess et al. (1998) are very similar. The overall grammatical distinction is made in both results, and internal semantic relationships are also present.

Grammatical information seems to be carried in HAL's word vectors. This can be seen for grammatical categories that vary greatly in their syntactic role (nouns, verbs, prepositions, and determiners). It can also be seen for more subtle distinctions within grammatical classes (determin-

ers, different types of verbs). The analyses presented here (and elsewhere; see Burgess & Lund, 1997a) suggest that HAL's word vectors carry important grammatical-class information and semantic information (also see Finch & Chater, 1992). This grammatical-class information appears to be sufficient to make coarsely coded distinctions (nouns, verbs, determiners) and more finely coded, within-class discriminations (Burgess & Lund, 1997a).

DISCUSSION

The HAL model offers an account of how simple episodic associations (defined here as local co-occurrences) are transduced into categorical knowledge. Word-meaning vectors represent the contextual history of a word—in a sense, the word's learning history. This contextual history (referred to as global co-occurrence) was shown to be informative in modeling a variety of categorical and semantic effects. Two domains—semantic and grammatical categorization—were addressed. We saw that HAL can be used to replicate the semantic priming effect by use of the distance metric. One of the first HAL results used the semantic distance metric to analyze two different sets of prime–target pairs from two different experiments. The analysis of Shelton and Martin's (1992) stimuli demonstrated that there was an asymmetry in the semantic relatedness that might have resulted in an overestimation of the effect of word association over semantic relatedness in determining the semantic priming effect. Although it is a commonly held belief that temporal contiguity is closely related to human word associations (Miller, 1969; Spence & Owens, 1990), Lund et al. (1996) found that the correlation between word-association ranking and (local) word co-occurrence was a product of the contextual similarity of the words. Words that are not similar to begin with did not show a correlation between association ranking and co-occurrence. In HAL, an association is probably best characterized as the local co-occurrence of items. However, it is the global co-occurrence (contextual similarity) that is most informative in modeling these semantic effects. Although there is evidence that suggests that semantic similarity does not produce automatic bottom-up priming (Lupker, 1984; Shelton & Martin, 1992), recent results with visual-neglect patients (Cushman et al., 1993) and with the same procedures used by Shelton and Martin (i.e., McRae & Boisvert, in press) converge on the conclusion that, with sufficiently similar word pairs, retrieval is rapid and automatic.

The priming results suggest that there is a meaningful organization of the spaces around a word. Neighborhoods of words close to a target word in the hyperspace mimic what could be viewed as a connotative definition. For example, the close neighbors to *beatles* reveal that the word has relationships with *original* and *band* and to the concepts *song* and *album*. More impressively, human subjects are able to use the words in the neighborhoods to converge on the word that produced the neighborhood in the first place. An important limitation of these neighborhoods, however, is that they reflect more the contextual connotation of meaning. Although a neighborhood may possess synonyms of a word, most neighbors do not suggest any kind of denotative definition. The neighborhoods are probably best thought of as a set of contextual cues that are useful in memory retrieval.

HAL's word vectors were also shown to be able to provide for the semantic categorization of a set of nouns (Figure 2), again with a result similar to that of a human faced with the same task. HAL's categorization ability was not limited to simple nouns, however. The vector representations encode information that also provides for grammatical categorization. In Figure 3, it is clear that nouns, verbs, and determiners occupy different parts of the high-dimensional space. Also, nouns representing common objects were segregated from the category of proper nouns corresponding to geographic locations, replicating the effect seen in Figure 2. More subtle differences within the class of determiners was also something that the vector representations were sensitive to (Figure 4). These grammatical and semantic categorical results suggest that word representations can encode what are typically considered different types of categorical information. From the perspective of a global co-occurrence model, context provides the basis that obviates the need for representational modularity. The idea is controversial and invites a rethinking of the nature of similarity.

Rethinking Similarity

Most notions of similarity hinge on the idea that concepts share some set of features (Komatsu, 1992). A car and a truck are similar because they both have tires, can be steered, transport people and things from place to place, and so on. These shared features provide the mechanism for similarity judgments, word recognition, or use by higher level comprehension systems. The notion of similarity in the HAL model is rather different. Concepts are similar because they occur in similar contexts. This is why, in our more recent work, we have been reluctant to refer to HAL as a semantic model—its representations encode meaning at a broader level of meaning. Meaning is a function of the contexts that words occur in, and, as a result, it offers an opportunity to consolidate, into one representational model, a diverse range of cognitive effects.

Using input units (words in the case of HAL) as the "features" by which meaning vectors are formed provides a clear grounding between symbol and environment. Berwick (1989) described the selection of labels for semantic or primitive features of meaning as a "hazardous game" (p. 95). The selection of any set of meaning elements is theoretically presumptive. Osgood et al.'s (1957) original approach of using the semantic differential required a commitment to a set of adjective pairs that delineates the semantic features. Global co-occurrence offers a partial solution to this problem. There is still a commitment to a set of meaning units, but the meaning units are, in the case of HAL, simply the input units. A similar co-occurrence approach has been successful in deal-

ing with the speech segmentation problem using phonetic segments (Cairns, Shillcock, Chater, & Levy, 1997). Thus, it becomes increasingly plausible to imagine an even more complete model of word meaning in which a simple learning algorithm could be used to develop different levels of representations, all capitalizing on the contexts in which words or phonemes are found.

Another important departure from the traditional view of similarity involves the relationship between nouns and verbs in sentences. In another paper (Burgess & Lund, 1997a), we discuss how the context distances between nouns and verbs (e.g., "the cop arrested ...") predicts whether or not human subjects will experience a processing load with syntactically ambiguous sentences. Using the traditional view of similarity, it is difficult to discuss how the concept *cop* and the concept *arrested* are similar. Clearly, there are important relationships between these two words: Arresting is part of the job of a cop, and cops are much more likely to arrest someone than to get arrested. But *cop* and *arrested* are not similar in the usual sense. How the entity *cop* and the action depicted by the verb *arrested* are similar is a function of context. Global co-occurrence is a process of concept acquisition that provides for the encoding of context in a straightforward fashion in a high-dimensional memory space. Furthermore, global co-occurrence allows for the encoding of concepts for which it is typically very difficult to imagine a set of features (abstract words, determiners, proper names). This contextual basis underlying the HAL model and the completely transparent way in which HAL accomplishes this are what provide the motivation to gravitate away from more limited approaches of thinking about conceptual features.

Conclusions

The HAL model uses a simple learning mechanism—global co-occurrence—to encode experience into high-dimensional representations of meaning. Although the input to the model is a large corpus of text, it is easy to imagine extending the basic notion of global co-occurrence to a more complete and realistic set of environmental contingencies. HAL has been useful as a model of word meaning and of the relationships between small numbers of words. A similar model by Landauer and Dumais (1997; also see Foltz, 1996), latent semantic analysis, has proven effective in capturing meaning at the sentential and discourse level. These models make a minimum of assumptions in how behavior is transduced into high-dimensional representations that are, in a very real sense, a history of learning. As a result, high-dimensional memory models provide an account of how simple associations become the semantic and grammatical building blocks of language.

REFERENCES

AU, T. K. (1986). A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory & Language*, **25**, 104-122.

BERWICK, R. C. (1989). Learning word meanings from examples. In D. L. Waltz (Ed.), *Semantic structures: Advances in natural language processing* (pp. 89-124). Hillsdale, NJ: Erlbaum.

BURGESS, C., LIVESAY, K., & LUND, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, **25**, 211-257.

BURGESS, C., & LUND, K. (1997a). Parsing constraints and high-dimensional semantic space. *Language & Cognitive Processes*, **12**, 177-210.

BURGESS, C., & LUND, K. (1997b). Representing abstract words and emotional connotation in high-dimensional memory space. *Proceedings of the Cognitive Science Society* (pp. 61-66). Hillsdale, NJ: Erlbaum.

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.

CHIARELLO, C., BURGESS, C., RICHARDS, L., & POLLOCK, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain & Language*, **38**, 75-104.

CUSHMAN, L., BURGESS, C., & MAXFIELD, L. (1993, February). *Semantic priming effects in patients with left neglect.* Paper presented at the meeting of the International Neuropsychological Society, Galveston, TX.

DEESE, J. (1965). *The structure of associations in language and thought* (pp. 97-119). Baltimore: Johns Hopkins University Press.

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

ERVIN, S. M. (1963). Correlates of associative frequency. *Journal of Verbal Learning & Verbal Behavior*, **1**, 422-431.

ERVIN-TRIPP, S. M. (1970). Substitution, context, and association. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 383-395). New York: Academic Press.

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: Erlbaum.

FISCHLER, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, **5**, 335-339.

FOLTZ, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, **28**, 197-202.

GARNHAM, A. (1996). The other side of mental models: Theories of language comprehension. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 35-52). Hove, U.K.: Psychology Press.

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

JOHNSON-LAIRD, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

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

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 Bulletin*, **104**, 211-240.

LUND, K., & BURGESS, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, **28**, 203-208.

LUND, K., & BURGESS, C. (1997). *Recurrent neural networks and global co-occurrence models: Developing contextual representations of word-meaning.* Paper presented at the NIPS*97 (Neural Information Processing Systems) Neural Models of Concept Learning Postconference Workshop, Breckenridge, CO.

LUND, K., BURGESS, C., & ATCHLEY, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the Cognitive Science Society* (pp. 660-665). Hillsdale, NJ: Erlbaum.

LUND, K., BURGESS, C., & AUDET, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. *Proceedings of the Cognitive Science Society* (pp. 603-608). Hillsdale, NJ: Erlbaum.

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

MCRAE, K., & BOISVERT, S. (in press). Automatic semantic similarity

priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

MILLER, G. (1969). The organization of lexical memory: Are word associations sufficient? In G. A. Talland & N. C. Waugh (Eds.), *The pathology of memory* (pp. 223-237). New York: Academic Press.

MORRIS, W. (ED.) (1971). *The American Heritage dictionary of the English language*. Boston: American Heritage.

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: Erlbaum.

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

OGDEN, C. K., & RICHARDS, I. A. (1923). *The meaning of meaning*. New York: Harcourt, Brace.

OSGOOD, C. E. (1971). Exploration in semantic space: A personal diary. *Journal of Social Issues*, **27**, 5-64.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

PALERMO, D. S., & JENKINS, J. J. (1964). *Word association norms grade school through college*. Minneapolis: University of Minnesota Press.

RIPS, L. J., SHOBEN, E. J., & SMITH, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning & Verbal Behavior*, **12**, 1-20.

SHELTON, J. R., & MARTIN, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1191-1210.

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

NOTES

1. I do not want to leave the impression that there is just one high-dimensional computational memory model. Kevin Lund and I started working on the HAL (Hyperspace Analogue to Language) model in 1992. The first conference presentation discussing HAL was in 1994, and the first publication was 1995. The work of Landauer and Dumais (1997, and earlier work; also see Foltz, 1996) is very closely related to our model. In fact, they are very close cousins—at a general level, the basic approach is the same. They differ in a number of implementational aspects, some of which may suggest important distinctions. Also, Nick Chater's (Finch & Chater, 1992) work takes a similar approach, although Chater is more reluctant to see these models as psychological models.

2. The process of frequency normalization and computing the distance metric are intertwined. The distance metric in HAL is referred to as *Riverside context units*. This is an arbitrary but frequency normalized Euclidian distance metric. The first step in normalizing a vector is to compute its magnitude (sum the squares of the elements, take the square root of that, and divide by the number of elements). The magnitude is then divided by 666.0, and each vector element is divided by the resulting number.

3. Visual inspection of the MDS presentations in this paper appears to show a robust separation of the various word groups. However, it is important to determine whether these categorizations are clearly distinguished in the high-dimensional space. Our approach to this is to use an analysis of variance that compares the intragroup distances with the intergroup distances. This is accomplished by calculating all combinations of item-pair distances within a group and comparing them with all combinations of item-pair distances in the other groups. In all MDS presentations shown in this paper, these analyses were computed, and all differences discussed were reliable.

(Manuscript received January 16, 1998;
revision accepted for publication March 5, 1998.)