Behavior Research Methods, Instruments, & Computers
1996, 28 (2), 203-208

Producing high-dimensional semantic spaces from lexical co-occurrence

KEVIN LUND and CURT BURGESS
University of California, Riverside, California

A procedure that processes a corpus of text and produces numeric vectors containing information about its meanings for each word is presented. This procedure is applied to a large corpus of natural language text taken from Usenet, and the resulting vectors are examined to determine what information is contained within them. These vectors provide the coordinates in a high-dimensional space in which word relationships can be analyzed. Analyses of both vector similarity and multidimensional scaling demonstrate that there is significant semantic information carried in the vectors. A comparison of vector similarity with human reaction times in a single-word priming experiment is presented. These vectors provide the basis for a representational model of semantic memory, hyperspace analogue to language (HAL).

Although there is a lack of agreement on exactly what semantic memory in humans is, some aspects of it are fairly noncontroversial. It is generally understood to be a memory of the meanings of things more than of their associations (the associations being maintained, perhaps, by more episodic memories). Human semantic memories are presumably constructed through experience with the world; as concepts are encountered, information about their meanings is accumulated. This paper will present and examine a method for creating a simulation that exhibits some of the characteristics of a human semantic memory, a simulation that develops through the analysis of human experience with the world in the form of natural language text.

Lexical co-occurrence has been established as a useful basis for the construction of semantic spaces (Burgess & Cottrell, 1995; Burgess & Lund, 1995b; Burgess & Lund, in press; Lund & Burgess, in press; Lund, Burgess, & Atchley, 1995). A semantic space is a space, often with a large number of dimensions, in which words or concepts are represented by points; the position of each such point along each axis is somehow related to the meaning of the word (Osgood, Suci, & Tannenbaum, 1957). Semantic spaces can be useful for examining the relationships between the words or concepts within them because, once the space is built, relationships can be quantified by applying distance metrics to the points within the space.

Semantic spaces are traditionally constructed by first defining the meanings of a set of axes and then gathering information from human subjects to determine where each word in question should fall on each axis. For example, a mouse might be placed near one end of a "size" axis, whereas a mountain would be closer to the opposite end (Osgood et al., 1957). There are both theoretical and practical problems with this approach. The theoretical problem is that the experimenter must choose a set of axes and hope both that they are sufficient to represent the desired level of detail in the space being built and that human judges will be able to accurately determine where stimuli should be placed along them. A more practical issue is that this is a tedious procedure: the number of subject judgments required is proportional to the number of axes desired multiplied by the number of items to be placed in the space. There are other approaches for determining semantic characteristics of words that involve judgments or ratings about the nature of interword relationships, but these still entail large numbers of human judgments to produce the semantics of relatively few items (Burgess & Lund, 1994; McRae, de Sa, & Seidenberg, 1993).

The use of lexical co-occurrence to construct semantic spaces addresses both of these problems. No explicit human judgments are required, and the choice of axes is, if not principled, at least no longer arbitrary. This paper presents a procedure by which high-dimensional semantic spaces may be constructed, in an automated fashion, from bodies of text and includes experiments illustrating how these spaces can model human concept similarity. This procedure underlies the development of our representational model of semantic memory, called hyperspace analogue to language (HAL). The procedure requires no explicit human judgments concerning word meanings outside of those implicitly expressed by the authors of the text being analyzed. The experiments presented here all use the same basic methodology to form vector representations of word meanings. This is the methodology we have used in a variety of experiments (Burgess & Lund, 1995b, in press; Lund & Burgess, in press; Lund et al., 1995) and is similar to that of Landauer and Dumais (1994), Schütze (1992), and Schvaneveldt (1990).

This research was supported by an NSF Presidential Faculty Fellow award (SBR-9453406) to C.B. Address correspondence to either author at the Psychology Department, 1419 Life Sciences Bldg., University of California, Riverside, CA 92521-0426 (e-mail: kevin@locutus.ucr.edu or curt@cassandra.ucr.edu).

BASIC METHODOLOGY

In this procedure, a "window," representing a span of words, is passed over the corpus being analyzed. The width of this window can be varied, and in Experiment 3 the effects of differing window sizes on model performance will be examined.

Words within this window are recorded as co-occurring with a strength inversely proportional to the number of other words separating them within the window. For instance, in the preceding sentence, the words "inversely" and "proportional" would receive a maximum co-occurrence value, while "inversely" and "separating" would be considered to co-occur more weakly (if the window was even wide enough to include them both).

By moving this window over the source corpus in one-word increments and recording, at every window movement, the co-occurrence values of the words within it, a co-occurrence matrix can be formed. This matrix has, as axes, the entire vocabulary under consideration, such that each cell of the matrix represents the summed co-occurrence counts for a single word pair. ("Word pair," in this discussion, is direction sensitive. Counts for the sequence "*xy*" and counts for the sequence "*yx*" are in different cells.)

This process produces a matrix in which, for every word in the target vocabulary, there is both a row and a column containing relevant values (for instance, the row may contain co-occurrence information for words appearing before the word under consideration, while the column contains co-occurrence information for words following it). This row/column pair may be concatenated so that, given an $n \times n$ co-occurrence matrix, a co-occurrence vector of length $2n$ is available. This vector of $2n$ length can be conceptualized as representing a word in $2n$ high-dimensional space. This can result in a very long vector, however. We have found that the effects reported in this paper rely on only the 100 to 200 most variant vector elements. Further reduction may be achieved, at the expense of computational complexity, by retaining some relatively small number of principal components of the co-occurrence matrix. Table 1 shows an example matrix computed for the sentence "the horse raced past the barn fell," using a window width of five words.

The corpus that we analyzed to produce the matrices examined in the experiments presented here consists of approximately 160 million words of text taken from Usenet newsgroups. The text was gathered during February of 1995 from all Usenet newsgroups carrying text. Usenet was chosen as a source of text for three main reasons. First, a virtually limitless supply of text is available; during the collection period, roughly 10 million words of new text were available each day. Second, it covers a very broad range of topics, which leads to a large range of potential word interactions. Third, the text is conversational and noisy, much like spoken language. No other available corpus offers these three advantages.

Table 1
Example Matrix for "The Horse Raced Past the Barn Fell"
(Computed for Window Width of Five Words)

	barn	fell	horse	past	raced	the
<PERIOD>	4	5	0	2	1	3
barn	0	0	2	4	3	6
fell	5	0	1	3	2	4
horse	0	0	0	0	0	5
past	0	0	4	0	5	3
raced	0	0	5	0	0	4
the	0	0	3	5	4	2

Once the matrices are constructed, similarity measurements can be applied to word vectors; this, we hoped, would yield a measure of semantic similarity between any desired pair of words. The distance metrics used in the following experiments all come from the Minkowski family of distance metrics, which include the familiar Euclidean (if $r = 2$) and city-block (if $r = 1$) metrics:

$$\text{distance} = \sqrt[r]{\sum(|x_i - y_i|)^r}. \tag{1}$$

Because these metrics are sensitive to vector magnitude, vectors were normalized to a constant length before the distance metrics were applied.

In Experiment 1, the distances between word vectors were examined to determine whether or not similarity in word meaning corresponded to similarity in patterns of vector elements. In Experiment 2, an attempt was made to demonstrate that the relationships found in Experiment 1 could be used in categorization and classification tasks. In Experiment 3, the effects of varying model parameters (window size and the distance metric) were explored and the model's performance on a more quantitative level as it relates to human performance was examined.

EXPERIMENT 1

If the procedure outlined above succeeds in capturing information about word meaning in the vectors developed, this should be apparent by examining similarities between vectors. Vectors representing words with similar meanings should themselves be similar, and, likewise, vectors for dissimilar words should have significant differences. See Figure 1 for examples of 25-element co-occurrence vectors. The vector comparison procedure used in this experiment will be the Euclidean distance between vectors; this corresponds to a Minkowski metric with $r = 2.0$.

Rather than select word pairs manually and examine the distances between them, we chose to examine, for each target word, the distances from that word of all other words in the vocabulary. These could then be ranked by distance to reveal the closest "neighbors" in the co-occurrence space to the target word. If the closest of these neighbors were words with meanings similar to that of the target word, we could conclude that the co-occurrence tabulation procedure was successful in evaluating relative word meanings.

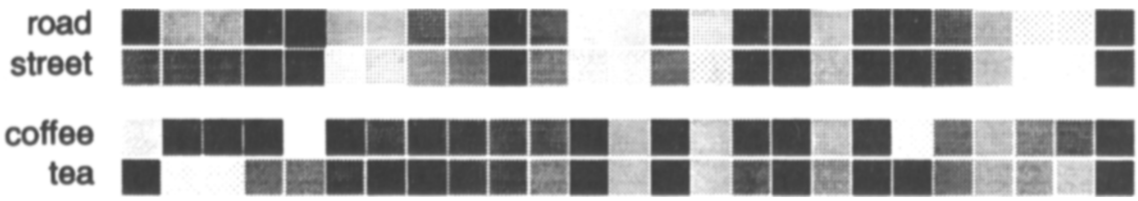

Figure 1. Gray-scaled 25-element co-occurrence vectors.

Method

A co-occurrence matrix was computed using a window size of 10 words. The input corpus was 160 million words of text from Usenet news groups. Any token appearing at least 50 times within the corpus was tracked as a vocabulary item, resulting in a vocabulary of roughly 70 thousand "words." Twenty target words were selected at random from the middle frequency range of the words appearing in the input corpus; limiting the frequency range in this manner eliminated both common function words and rare technical terms.

For each target word, a Euclidean distance was computed from the word to each vocabulary item, resulting in a set of 70 thousand distances. These were then sorted, and the neighbors with the smallest distances were examined. A sample of the results is shown in Table 2.

Results and Discussion

Examination of Table 2 shows that near neighbors of words in the co-occurrence space share aspects of meaning. These relationships appear to be both semantic (*jugs–cans*, *cardboard–plastic*) and associative (*lipstick–lace*, *monopoly–threat*). The results suggest that the high-dimensional neighborhood surrounding each word is something akin to a semantic field.

EXPERIMENT 2

In this experiment we sought to determine whether or not the co-occurrence matrix-construction procedure would result in word vectors that carried categorical information. Multidimensional scaling has been shown to be useful in finding structure in data that can be represented by item similarity evaluations (Shepard, Romney, & Nerlove, 1972). In this experiment, multidimensional scaling was applied to a set of co-occurrence vectors to determine whether or not intervector distances could provide a credible substitute for human similarity judgments.

Method

Words representing three categories (animal names, body parts, and geographical locations) were chosen, and word vectors for these words were extracted from a co-occurrence matrix that had been constructed using a co-occurrence window width of 10 words. Item similarities were computed using a Minkowski r metric with $r = 2$ (see Equation 1). A multidimensional scaling was performed on this similarity data to produce a two-dimensional solution.

Table 2
Five Nearest Neighbors for Target Words
From Experiment 1 (*n*1 ... *n*5)

Target	*n*1	*n*2	*n*3	*n*4	*n*5
jugs	juice	butter	vinegar	bottles	cans
leningrad	rome	iran	dresden	azerbaijan	tibet
lipstick	lace	pink	cream	purple	soft
triumph	beauty	prime	grand	former	rolling
cardboard	plastic	rubber	glass	thin	tiny
monopoly	threat	huge	moral	gun	large

Results and Discussion

Figure 2 shows the obtained solution with lines added to help clarify the delineation of categories. The geographical locations, being very unlike either body parts or animals, are clearly represented as a distinct group. The body parts and animal types are also quite well separated, with only the body part "tooth" intruding into the cluster of animal names. Intuitively, "tooth" is a particularly salient body part for animals.

This result validates the basic methodology. No explicit human judgments about item similarity were used in this procedure, and yet the simulation results in a plausible categorization of three sets of items. This demonstrates that the co-occurrence procedure used was successful in extracting general semantic information from the corpus.

EXPERIMENT 3

This experiment expanded on the results of Experiment 2 by comparing vector similarities with reaction times from a lexical priming study across a range of methodological parameters (manipulating co-occurrence window size and distance metric). Word similarity has been established as an important factor in subjects' reaction times to prime–target pairs (Chiarello, Burgess, Richards, & Pollock, 1990; Lund et al., 1995; Neely, 1977), so an analysis of the relationship between vector similarities for word pairs and lexical decision times to the targets of the same word pairs should help to illuminate the relationship between word meanings and co-occurrence vectors.

Method

Related prime-target pairs from Chiarello et al. (1990) were used, along with unrelated word pairs as controls, in a lexical decision task. Subjects were presented with a fixation point, followed by the prime word; after a 250-msec SOA, the target word (or a nonword) was presented and the subject's decision time was recorded.

Vector similarities were computed for these word pairs. Six co-occurrence matrices were analyzed. The matrices varied in the window width used to construct them (1-, 2-, 4-, 6-, 8-, and 10-word windows). Distances were computed using each of three r metrics (rs = 1, 1.5, and 2). Correlations were calculated between vector similarities for word pairs and reaction times for those same pairs.

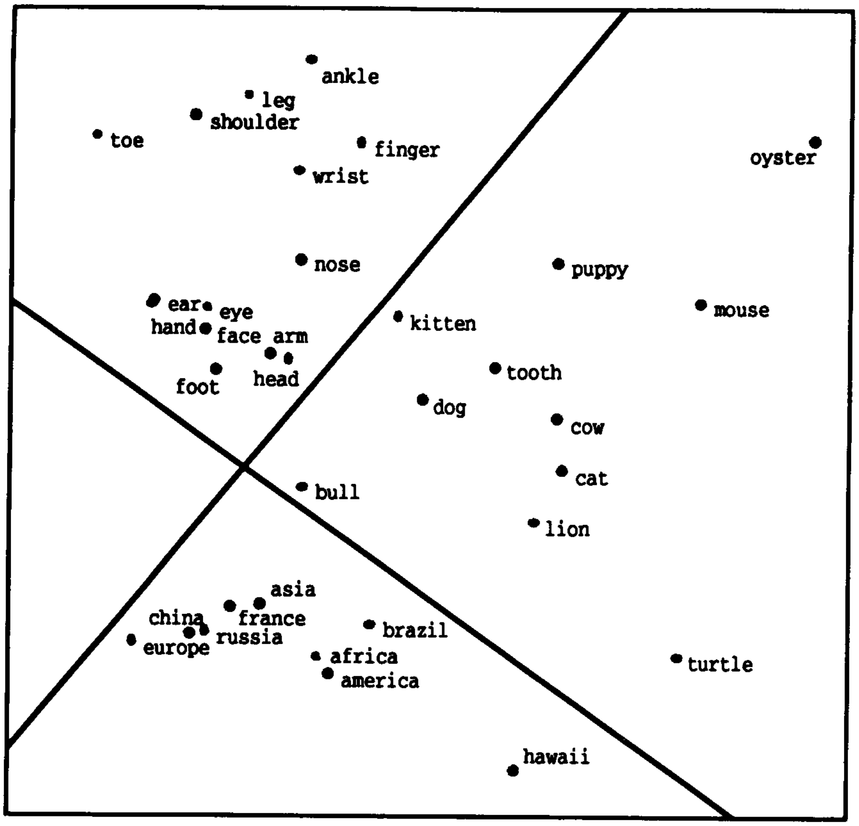

Figure 2. Multidimensional scaling of co-occurrence vectors.

Results and Discussion

Figure 3 shows the correlations between semantic distance and reaction time (all correlations significant, $p <$ .0001). Generally, performance improved as r decreased; performance also varied considerably with window size, with a peak at width 8. Optimal performance was obtained using an r of 1.0. This result is consistent with Shepard's (1980) proposal that r values close to 1 might be best suited for analyses in the semantic domain.

This experiment demonstrates a sizable correlation between vector similarity and basic cognitive effects. It also helps to establish desirable parameters for the co-occurrence analysis process. The correlations obtained in this experiment between semantic distance and reaction time are similar to the results reported by Fischler (1977). He found a correlation of 0.31 between human similarity estimates and the single-word priming effect in his experiment.

GENERAL DISCUSSION

This series of experiments has described a methodology capable of capturing information about word meanings through the unsupervised analysis of text. This promises to be a generally useful technique for the analysis of word semantics, as well as providing an alternative to the traditional methods of generating feature vectors for words.

The use of lexical co-occurrences in obtaining semantic information requires further discussion. Lexical co-occurrence has been shown to correlate with human association norms (Spence & Owen, 1990). In Experiment 1, nearest neighbors included items that seemed to be both semantically related as well as associatively related. Whether or not related word pairs are related associatively, their semantic relationship, or some combination of the two, has often, if not typically, been confounded (Chiarello et al., 1990). There is evidence that the vectors produced by this procedure are more semantic, that is, similarity based, than associative in nature. In a simulation of a variety of semantic and associative priming results, these semantic vectors provided a robust relatedness effect (semantic distance of related pairs < semantic distance of unrelated pairs) for items that were semantically, but not associatively, related, for example, *table–bed.* Items that were associatively, but not semantically, related did not show a reliable relatedness effect, for example, *coffee–cup* (Lund et al., 1995).

The implication of these results is important. The vectors generated by HAL function semantically. Although their genesis resides in the first-order association of lexical co-occurrence, the vectors do not ultimately rely on the associative components to provide the semantic effect. This claim is illustrated by a comparison of the word vectors shown in Figure 1. The concepts of *coffee* and *tea* are similar and strongly associated. They tend to co-occur

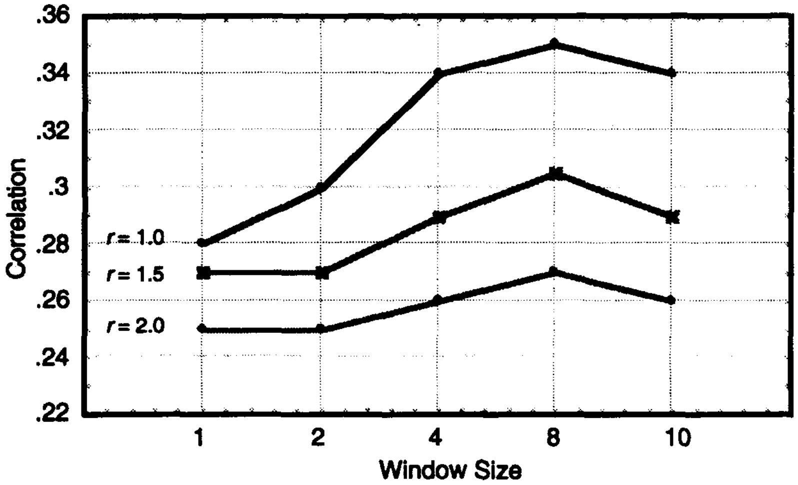

Figure 3. Correlations between vector distances and human reaction times for three distance metrics.

in natural language, and their similarity can be seen in their vector representations. These can be contrasted to *road* and *street*. Again, although these are two concepts that are highly similar, they do not tend to co-occur in usage, yet their vector representations are very similar.

HAL acquires word meanings as a function of keeping track of how words are used in context by using the moving window and weighting co-occurrence distance. Over millions of words of experience, the matrix that develops carries the history of this contextual experience. Similar representations *tend* to be words that can be substituted for each other in context. For example, such semantically similar items as *bed* and *table* are relatively interchangeable in Sentences 1a and 1b. Contrast this with such items as *cradle* and *baby*, which are only associatively related in Sentences 2a and 2b:

1a. The child slept on the bed.
1b. The child slept on the table.

2a. The child slept in the cradle.
2b. The child slept in the baby.

The semantic vectors are representations that are essentially measures of context; that lexical co-occurrence is an important component in the acquisition process becomes independent of the eventual generalizations that the semantic vectors represent. It has been a long-held belief that word relations evolve through both co-occurrence in language as well as the substitutability of words in contexts (Ervin-Tripp, 1970). These vectors capture the substitutability, the semantic similarity, of word relations.

In the present series of experiments, the representations that HAL develops have demonstrated some basic effects that would be required of microstructural semantics. In Experiment 1, the "nearest neighbors" demonstration showed that the semantic field surrounding the concept includes important aspects of a word's relationships to other words. Experiment 2 demonstrated that the vectors can be used to classify instances of superordinate categories. Finally, in Experiment 3, the semantic distances were shown to correlate with human reaction times in a lexical priming study.

In other experiments, semantic vectors generated by using HAL have accounted for a range of semantic and associative priming results using stimuli from various investigators (Lund et al., 1995). Semantic constraints used during higher level sentence comprehension and syntactic processing can be characterized by using these vectors (Burgess & Lund, 1995a). The semantic representations have also been used as the representational basis for developing a model of cerebral asymmetries in lexical/ semantic processing (Burgess & Lund, in press). This broad range of results suggests that these representations provide the basic representational microstructure that can be used to model various aspects of human semantic memory, to analyze stimuli, or to construct information-retrieval tools.

HAL represents a procedure for acquiring semantic representations in an unsupervised fashion in a noisy, conversation-like environment without the heavy preprocessing of text required by other automated systems (Armstrong, 1994; Zernik, 1991). The methodology easily exploits the regularities of language such that conceptual generalizations can be captured in a data matrix, making it straightforward to use in modeling human memory.

REFERENCES

Armstrong, S. (Ed.) (1994). *Using large corpora*. Cambridge, MA: MIT Press.

Burgess, C., & Cottrell, G. (1995). Using high-dimensional semantic spaces derived from large text corpora. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 13-14). Hillsdale, NJ: Erlbaum.

Burgess, C., & Lund, K. (1994). Multiple constraints in syntactic ambiguity resolution: A connectionist account of psycholinguistic data. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 90-95). Hillsdale, NJ: Erlbaum.

Burgess, C., & Lund, K. (1995a). *High-dimensional semantics from* *corpora and human syntactic processing constraints*. Paper presented at the 8th Annual CUNY Sentence Processing Conference, Tucson, AZ.

Burgess, C., & Lund, K. (1995b, November). *Hyperspace analogue to language (HAL): A general model of semantic representation*. Paper presented at the annual meeting of the Psychonomic Society, Los Angeles.

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: Erlbaum.

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.

Ervin-Tripp, S. M. (1970). Substitution, context, and association. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 383-467). New York: Academic Press.

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

Landauer, T. K., & Dumais, S. (1994, November). *Memory model reads encyclopedia, passes vocabulary test*. Paper presented at the annual meeting of the Psychonomic Society, St. Louis.

Lund, K., & Burgess, C. (in press). A general model of semantic representation (abstract). *Brain & Cognition*.

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

McRae, K., de Sa, V., & Seidenberg, M. S. (1993). *The role of correlated properties in accessing conceptual memory*. Unpublished manuscript.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, **106**, 226-254.

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

Schütze, H. (1992). *Dimensions of meaning*. Unpublished manuscript.

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

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, **210**, 390-398.

Shepard, R. N., Romney, A. K., & Nerlove, S. B. (Eds.) (1972). *Multidimension scaling: Theory and applications in the behavioral sciences*. New York and London: Seminar Press.

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

Zernik, U. (Ed.) (1991). *Lexical acquisition: Exploiting on-line resources to build a lexicon*. Hillsdale, NJ: Erlbaum.

(Manuscript received December 4, 1995;
accepted for publication December 19, 1995.)