Borrower: BUF

Lending String: *NAM,YSM,BNG,SYB,XBM

Patron: Rapaport, William

Journal Title: Basic and applied memory research /

Volume: 1 **Issue:**
Month/Year: 1996**Pages:** 105-126

Article Author:

Article Title: Landauer, TK; Dumais, ST; How come you know so much? From practical problem to new memory theory

Imprint: Mahwah, N.J. ; Lawrence Erlbaum Associat

ILL Number: 35020630

Call #: BF 371 B27 1996

Location:

ARIEL
Charge
Maxcost: $30.00IFM

Shipping Address:
University at Buffalo
BUF - BUFFALO HUB - BUFFALO
LAND DELIVERY

Fax:
Ariel: 128.205.111.1

SH

CHAPTER SEVEN

How Come You Know So Much? From Practical Problem to New Memory Theory

Thomas K. Landauer
Bellcore and the University of Colorado

Susan T. Dumais
Bellcore

When we talk about the relation between theory and practice, we usually assume that the science comes first and enables the solution of practical problems. Thus, we speak of "applied science" and have no equally common noun phrase for direct attack on practical problems by scientific means, although "engineering research" comes close. Yet many historians, philosophers, and commentators have concluded that the historical path is more often from practice to science than from science to practice (Kuhn, 1977; Mokyr, 1990; Petroski, 1982).Technology usually advances by incremental trial, error, decomposition, simulation, and improvement, with general principle discovery and scientific theory occasionally sprouting from the process and occasionally helping to solve future practical problems. The Wright brothers perfected their wing shapes by hundreds of trials in a wind tunnel, not by calculations from aerodynamic theory. Aerodynamic theory mostly came later, to explain why the wing shapes that Wilbur and Orville chose had adequate lift and stability and to suggest new versions that, even today, need wind-tunnel testing before major investment. Fisher invented analysis of variance to help agronomists select among seed varieties, not to help biologists—much less psychologists—prove theories. On the other hand, the needs of practical problems often drive, or at least stimulate, science and theory. Navigation motivated astronomy; artillery and commerce motivated geometry and physics; medicine motivates molecular biology. Scientists try to discover the biochemical, cellular, genetic, or physiological processes that account for disease states in order to support rational cures for well-known diseases more often than they take discoveries from pure science and seek diseases to which to apply them. Sometimes science is stimulated by failures of practice, by the appearance of unsuccessful or dangerous technology. The collapse of large numbers of early iron bridges was eventually *followed* by scientific investigation of the physical properties of iron beams (Petroski, 1982).

This volume's theme is theory in context. The offering that we bring is a case in which the partial solution of a practical problem in information retrieval has given rise to the germ of a theory that might resolve an empirical mystery about human memory. We start by describing the mystery, which in itself is related to a highly practical problem. We then review an apparently unrelated program of psychological engineering research, which in the end gave rise to a practically useful mathematical model and method. Finally, we show how this model can be viewed as a theory of certain aspects of human memory, report some evidence of its success as such a model, discuss how it might solve the original empirical problem, and propose tests to see whether its mechanisms should be incorporated into our general theories of memory.

THE MEMORY CONUNDRUM: CHILDREN LEARN VOCABULARY TOO FAST

The empirical problem is this: The average college graduate knows the meaning of about 100,000 distinct words. Many readers of this chapter may know twice that many. The way such numbers have been estimated is to choose words at random from a large dictionary, do some kind of test on a sample of people to see what proportion of the words they know, then reinflate. Several researchers have made such estimates (see Nagy & Herman, 1987). The varying totals they come up with are largely determined by the size of the dictionaries that they start with, and to some extent with the way in which they define words as being separate from each other. Here is one example of an estimation procedure. Moyer and Landauer (Landauer, 1986) sampled 1,000 words from *Webster's Third Unabridged Dictionary* and presented them to Stanford undergraduates along with a list of 30 common categories. If a student classified a word correctly and rated it familiar it was counted as known. Landauer then went through the dictionary and guessed how many of the words could have been gotten right by knowing some other morphologically related word, and adjusted the results accordingly. The resulting estimate was around 100,000 words. This is at the high end, but is roughly consistent with numbers from more careful studies in the literature when extrapolated to high-ability young adults.[1] It appears that

[1]Nagy and Anderson (1984), starting with a word list based on schoolbooks (Carroll, Davies, & Richman, 1971) and using a similar method, estimated 40,000 words for average high school seniors.

even this estimate may be somewhat l
words found in a daily newspaper d
names, some quite common (Walker &
Knowing 100,000 words by 20 year
of about 15 words a day from age 2 or
elementary and high school years has l
5,400 words per year (10 to 15 per da
twice as rapid gains as the average (N
Thus, normal schoolchildren appear c
per day over sustained periods.
Most words are learned by reading.
vocabulary is much smaller than writt
viduals hear in daily intercourse with
for less than one fifth their reading con
children spend more than a third of the
sets, and the vocabulary of television
little vocabulary is learned from direct
little time to it, and it produces meager
100 words a year could come from th
Estimates are that the average fifth-gr
day reading in school and another 15 ou
mail, and comic books (Anderson, Wil
Maruyama, 1990). If we assume 30 min
and 15 minutes per day for the rest of the
per day. Thus, while reading, kids ar
minute. Combining estimates of reader
& Anderson, 1985) with an average rea
(Anderson & Freebody, 1983; Carver, 1
that young readers encounter about thr
Thus, the opportunity is there to acqui
terrifically rapid rate of learning. Consic
ing speed. You'd have to give childre
definitions each day and expect them t
very brief study trial.[2] Never have w
classrooms, laboratories, or learning th
Word knowledge comes from readi
have tried to mimic the contextual lea
usually done by selecting nonsense or
grade-level vocabulary knowledge an

[2]Remarkably, Pressley, Ross, Levin, and Gh
definition pairs after only one key-word strategy
retention of 11 items.

even this estimate may be somewhat low, because as many as 60% of the words found in a daily newspaper do not occur in dictionaries—mostly names, some quite common (Walker & Amsler, 1986).

Knowing 100,000 words by 20 years of age means learning an average of about 15 words a day from age 2 onward. The rate of acquisition during elementary and high school years has been estimated at between 3,000 and 5,400 words per year (10 to 15 per day), with some years showing almost twice as rapid gains as the average (Nagy & Herman, 1987; Smith, 1941). Thus, normal schoolchildren appear capable of learning at least 20 words per day over sustained periods.

Most words are learned by reading. The proof is straightforward. Spoken vocabulary is much smaller than written vocabulary. The words that individuals hear in daily intercourse with family and friends probably account for less than one fifth their reading comprehension vocabulary. Most schoolchildren spend more than a third of their waking hours in front of television sets, and the vocabulary of television discourse is even more limited. Very little vocabulary is learned from direct instruction. Most schools devote very little time to it, and it produces meager results. Authorities guess that at best 100 words a year could come from this source (Durkin, 1979).

Estimates are that the average fifth-grade child spends about 15 minutes per day reading in school and another 15 out of school reading books, magazines, mail, and comic books (Anderson, Wilson, & Fielding, 1988; Taylor, Frye, & Maruyama, 1990). If we assume 30 minutes per day total for 150 school days and 15 minutes per day for the rest of the year, we get an average of 21 minutes per day. Thus, while reading, kids are learning about one new word per minute. Combining estimates of reader and text vocabularies (Nagy, Herman, & Anderson, 1985) with an average reading speed of 165 words per minute (Anderson & Freebody, 1983; Carver, 1990; Taylor et al., 1990), we can infer that young readers encounter about three not-yet-known words per minute. Thus, the opportunity is there to acquire the daily ration. However, this is a terrifically rapid rate of learning. Consider the necessary equivalent list-learning speed. You'd have to give children a list of 60 new words and their definitions each day and expect them to permanently retain 20 after a single very brief study trial.[2] Never have we seen such a learning rate in our classrooms, laboratories, or learning theory parameter fits.

Word knowledge comes from reading, but how? Several research groups have tried to mimic the contextual learning of words. The experiments are usually done by selecting nonsense or unknown words at the frontier of the grade-level vocabulary knowledge and embedding them in carefully con-

---

[2]Remarkably, Pressley, Ross, Levin, and Ghatala (1984) reported 51% learning of word-definition pairs after only one key-word strategy learning trial. However this was for short term retention of 11 items.

structed sentences or paragraphs that imply aspects of meaning for the words. The results are uniformly discouraging. For example, Jenkins, Stein, and Wysocki (1984) constructed paragraphs around 18 low-frequency words and had fifth graders read them up to 10 times each over several days. The chance of learning a new word on one reading, as measured by a forced choice definition test, was between .05 and .10. More naturalistic studies have used paragraphs from schoolbooks and measured the chance of a word moving from incorrect to correct on a later test as a result of one reading (Nagy et al., 1985). About one out of 20 words makes the jump. Thus, experimental attempts to induce vocabulary acquisition through reading have achieved less than one sixth the natural rate when trying to simulate real reading, and less than one third even when explicitly trying to outdo nature.

So what's going on? How is it that children learn words from context at a rate much greater than we can get them to intentionally? The explanation we will offer did not occur to us until after an entirely independent research effort on information retrieval, so we will tell it in that order as well.

The Engineering Problem

In the early 1980s, four psychologists at Bell Labs were working more or less independently on techniques by which users could communicate with computers. George Furnas was collecting names for categories for an on-line classified ad prototype, Louis Gomez was creating indexes for a recipe file to use in experiments, Sue Dumais and Tom Landauer were having students name statistically derived clusters of yellow page headings, and Landauer and Kathleen Galotti were trying to find better names for text editor commands. Everyone found that nobody agreed on what to call anything. There was no consensual "natural" name for an editing command, no consensual title for a classified or yellow pages ad category, little overlap in key words assigned by cooks to the same recipe. Frustrated in our hopes to cure computer usability problems by finding natural, easy to learn terminology, we decided to study the problem before solving it. We pooled our data, gathered more from others, and did elaborate statistical analyses and simulations, leading to an incredibly long and detailed paper in the *Bell System Technical Journal* (Furnas, Landauer, Gomez, & Dumais, 1983a, 1987). In it we declared the opening of a new field of research that we called "Statistical Semantics," of which that article was the first and, as far we know, last example.

The central finding of all this effort was that although some linguists (e.g., Clark, 1987; Pinker, 1994) will tell you that there is no such thing as a true synonym, any object that you ask people to name, especially information objects like advertised items or abstracts of documents, will be referred to by about 30 different terms. If you ask for preferred terms from each of 100 people, between them they will come up with 10 to 50. Each person will think of between three and seven, and there will be little overlap among them. The chance that two people will choose the same word as their most favored moniker is somewhere between 10% and 20%.

It occurred to us that the difficulties people encounter in trying to look things up in on-line databases, or for that matter in card catalogs or back-of-book indexes, might be due to this disagreement in verbal labeling. We learned that professional indexers were aware of this problem, knew that they were unreliable in assigning keywords, and that their clients were even less to be trusted. However, they had never done the type of psychological experiments that we had, or the kinds of simulations, and were unaware of just how severe the problem was. Indeed, the common approach to its solution had been, and still is, to define and enforce a standard vocabulary for a particular domain of knowledge (e.g., chemistry or medicine) and to try to train all indexers to apply the same words. It hasn't worked. Indexers still disagree about half the time, and the untrained users who actually want the materials are hopelessly prone to ignoring the controlled vocabulary in favor of words they can think of.

We went on to study—by both simulation and direct experiment—ways to overcome the synonymy problem. We discovered that if we actually collected all the words that anybody wanted to apply to a given abstract or command, things got much better. As we went from assigning just one keyword to an item to assigning an average of 30, the chances that a user's spontaneous entry would match a desired target increased from under 20% to almost 80%. We dignified this finding in a principle that we called "Unlimited Aliasing": If users want to call something by a particular term, let them.

Libraries and publishers had never dared give each item 30 index entries before, partly because in paper the bulk would be unwieldy, and partly because they feared that extra words would lead to unwanted ambiguity. It turned out that the fear of ambiguity was largely unwarranted. Additional index words did increase somewhat the number of irrelevant things pointed to, but not nearly as much as they improved the likelihood of finding something a searcher wanted. In part this is because words that are thought of later tend to be more specific. More important, terminology is many-to-one from terms to objects—because there are many different aspects of an object that a person may want to specify—but each term tends to have only one or a few meanings in any particular domain. People tend to notice when they get things they don't want, but don't know how many things they miss. As a result, there is a pervasive overemphasis on the ambiguity (false positive) problem at the expense of the more important recall (hit rate) problem in information retrieval.

The next problem was to find a way to collect all the terms that were needed. The most effective method so far devised is "Adaptive Indexing," a technique invented by George Furnas (Furnas, 1985). It is well illustrated by an experimental prototype that he built for the on-line directory of campus services at the University of Texas. When a user typed in a keyword such as "Reproduction," the machine in its original form would come back with the response " 'Reproduction' not known." The same negative response was provoked by the keywords "Copying" and "Xerox." The frustrated user would ask around and discover that the desired department was actually called "Reprographics." She would type "Reprographics," and the machine would say, "Reprographics department does reproduction, Xeroxing, and copying. Tel. No. NNNN." Before the user could quit, the machine would ask, "Do you think the words 'reproduction,' 'Xerox,' and 'copying,' should be added to the index terms for the reprographics department?" With user concurrence, they are. The next time this user or anyone else in the community types in "Reproduction," the system will return the "Reprographics Department" among its possible choices. With repeated uses, the system acquires just those terms that most people want to apply most often to just those things that they most often have trouble finding. After a while the system will have collected a tally of how often each entered word was satisfied by particular answers. Then the system might return in response to the query Reproduction: Reprographics Department—60, Model Shop—5, Health Clinic—1. The user then chooses the most fitting option for his or her needs, with the possibility of asking for more information about each. In Furnas' field trial the system improved the probability of getting a correct answer by 50% after only a few hundred uses.

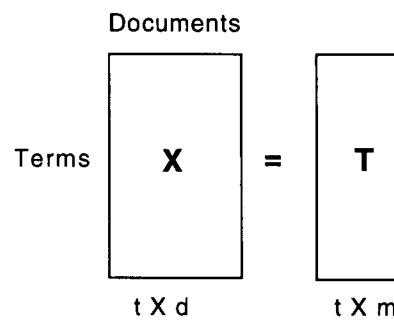
Unfortunately, in very large, rapidly changing collections, such as the medical literature, it is not satisfactory to wait for the user population to provide the necessary aliases. Many important searches may be the first ever for a particular document. Therefore, we wanted an automatic analysis method that could do some of the same job. We needed a way to discover and represent the relationship between words and the textual objects to which they might refer. The state-of-the-art technique in machine information retrieval is called the "vector method." In this approach, documents or, more properly, document surrogates such as titles or abstracts are represented as an unordered set of the words that they contain. A collection of documents is then represented as a large matrix in which each word contained in any document (absent a few hundred rare or too frequently occurring words) is a row or dimension, and each document is a column, the cells containing the number of times that a particular word occurs in a particular document. (In actual application, some transform is usually applied to the cell entries to weight most heavily those that carry the most information about which documents they are in.) A user or searcher query is construed to be the same sort of vector as a document and is compared by some pattern matching metric to each of the documents in the collection, and the system returns a list in order of the degree of match. (The degree of match is usually measured

by a cosine between the document
that measure works best, although v
torily explained.)

Unfortunately, this method does r
problem. It treats each term as totall
a separate dimension. A one word q
practice—will have a zero cosine m
contain it, even though the docum
terminology that has a very close m

What we wanted was a method tl
lying structure in a word-by-docum
cussing similar topics would have s
long as each word is its own separa
analysis that reduces the dimensior
appropriate and computationally tra

We chose a linear decomposition
position, or SVD. This is a form of
mathematical generalization of whic
SVD a rectangular matrix is decom
matrices (see Fig. 7.1). One compon
a vector of derived orthogonal facto
entities in the same way, and the thir



Singular Value Decomposition of

T has orthogonal unit-len
D has orthogonal unit-len
S is the diagonal matrix c
t is the number of of rows
d is the number of colum
m is the rank of X (<= mi

FIG. 7.1. Schematic of the singu
rectangular term by document matrix
three matrices each with linearly ind

built for the on-line directory of campus
When a user typed in a keyword such
s original form would come back with
own." The same negative response was
' and "Xerox." The frustrated user would
lesired department was actually called
eprographics," and the machine would
s reproduction, Xeroxing, and copying.
uld quit, the machine would ask, "Do
Xerox,' and 'copying,' should be added
ics department?" With user concurrence,
anyone else in the community types in
turn the "Reprographics Department"
epeated uses, the system acquires just
o apply most often to just those things
ding. After a while the system will have
ntered word was satisfied by particular
rn in response to the query Reproduc-
, Model Shop—5, Health Clinic—1. The
option for his or her needs, with the
nation about each. In Furnas' field trial
of getting a correct answer by 50% after

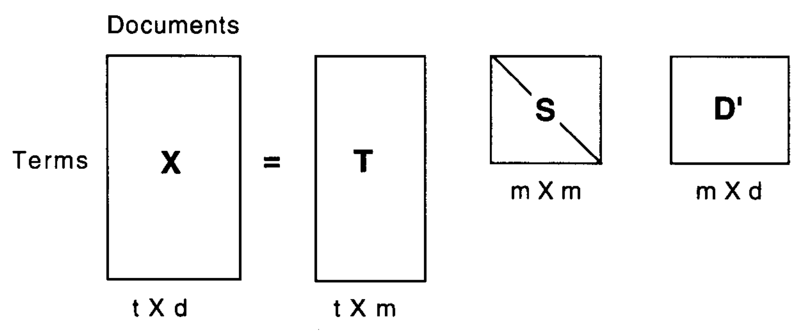
idly changing collections, such as the
ory to wait for the user population to
mportant searches may be the first ever
re, we wanted an automatic analysis
ame job. We needed a way to discover
veen words and the textual objects to
he-art technique in machine information
." In this approach, documents or, more
as titles or abstracts are represented as
hey contain. A collection of documents
in which each word contained in any
are or too frequently occurring words)
ument is a column, the cells containing
word occurs in a particular document.
rm is usually applied to the cell entries
arry the most information about which
searcher query is construed to be the
d is compared by some pattern matching
he collection, and the system returns a
The degree of match is usually measured

by a cosine between the document and query, for the simple reason that that measure works best, although why it does so has never been satisfactorily explained.)

Unfortunately, this method does nothing about the verbal disagreement problem. It treats each term as totally independent of every other, each as a separate dimension. A one word query—as the majority of queries are in practice—will have a zero cosine match with any document that does not contain it, even though the document may be highly pertinent and use terminology that has a very close meaning.

What we wanted was a method that would extract and represent underlying structure in a word-by-document matrix, so that words used in discussing similar topics would have similar vectors. This cannot happen as long as each word is its own separate dimension. What is needed is some analysis that reduces the dimensionality of the space and does so in an appropriate and computationally tractable way.

We chose a linear decomposition method called *singular value decomposition*, or SVD. This is a form of factor analysis, or more properly the mathematical generalization of which factor analysis is a special case. In SVD a rectangular matrix is decomposed into the product of three other matrices (see Fig. 7.1). One component matrix describes the row entities as a vector of derived orthogonal factor values, another describes the column entities in the same way, and the third is a diagonal matrix containing scaling



Singular Value Decomposition of the term by document matrix, X. Where:

T has orthogonal unit-length columns (T' T = I)
D has orthogonal unit-length columns (D' D = I)
S is the diagonal matrix of singular values
t is the number of of rows of X
d is the number of columns of X
m is the rank of X (<= min (t,d))

FIG. 7.1. Schematic of the singular value decomposition (SVD) of a rectangular term by document matrix. The original matrix is decomposed into three matrices each with linearly independent components.

values such that when the three components are multiplied, the original matrix is reconstructed. There is a mathematical proof that any matrix can be so decomposed perfectly using no more factors than the smallest dimension of the original matrix. When fewer than the necessary number of factors are used, the reconstructed matrix is a least-squares best fit.

SVD had several nice properties for our purpose. First, we could control the number of dimensions precisely, using as many as necessary to represent all the different word meanings in a domain but, presumably, not so many as to represent different words with similar usage as unrelated. At least that was the hope. By dropping the smallest dimensions, by hypothesis we reduce the influence of unimportant differences between words and between documents, such as which of two words of related meaning was used in a particular document.

How SVD/LSI Works

Just as in the straight vector method, a collection of documents is cast as a large matrix of words by segments of text (documents); the cells contain a weighted transform of the number of times a word occurs in a document.[3] The matrix is submitted to SVD. (Because of recent advances in sparse-matrix algorithms and computer power, collections on the order of 50,000 documents containing 70,000 useful word types can now be analyzed on popular workstations in a few hours.) The number of dimensions kept is usually determined empirically by trying a set of queries and seeing what gives the best results. For many purposes, 150–350 dimensions works well, with a gently peaked optimum. More than the optimum and SVD begins to approximate the original matrix too closely and lose the advantage of the reduced structure; fewer, and the representation lacks sufficient discrimination. [We make no attempt to rotate or interpret the dimensions; there is no need or point. They can be thought of simply as abstract dimensions of lexical usage. For more on all this, see Deerwester, Dumais, Furnas, Landauer, & Harshman (1990).]

Here is a small example that gives the flavor and demonstrates some of what the technique accomplishes. This example uses as document surrogates just the titles of nine technical memoranda produced one year on our floor at Bellcore. Five of the nine were about human computer interaction, and four about mathematical graph theory. The original matrix has nine columns, and we have given it 12 rows, each corresponding to a content word used in at least two of the titles. The titles, with the indexed terms italicized, are shown in Fig. 7.2a. The corresponding word-by-document matrix is shown in Fig.

[3]The usual transform we have applied, including for the analyses reported here, weights terms inversely with their entropy (i.e., -sum p log p over all documents) and cell entries as their logs.

(a)

Titles of Technical Memos

c1: *Human* machine *interface* for ABC comput

c2: A *survey* of *user* opinion of *computer syster*

c3: The *EPS user interface* mangement *system*

c4: *System* and *human system* engineering tes

c5: Relation of *user* perceived *response time* to

m1: The generation of random, binary, order

m2: The intersection *graph* of paths in *trees*

m3: *Graph minors* IV: Widths of *trees* and wel

m4: *Graph minors*: A *survey*

FIG. 7.2a. A sample dataset consis memoranda. Terms occurring in more two classes of documents—five about and four about mathematical graph th

(b)

X =

	c1	c2	c3	c4
human	1	0	0	1
interface	1	0	1	0
computer	1	1	0	0
user	0	1	1	0
system	0	1	1	2
response	0	1	0	0
time	0	1	0	0
EPS	0	0	1	1
survey	0	1	0	0
trees	0	0	0	0
graph	0	0	0	0
minors	0	0	0	0

FIG. 7.2b. This dataset can be describ matrix in which each cell entry indica occurs in a document.

7.2b. The linear decomposition is shov cross multiplication perfectly reconstru show a reduction to just two dimensi original matrix. This two dimensional geometrical representation of the dime The two-dimensional vector for each dc on a plane for each, as shown in the f points representing both terms and doc distances between terms and betweer

(a)

Titles of Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* mangement *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

FIG. 7.2a. A sample dataset consisting of the titles of nine technical memoranda. Terms occurring in more than one title are italicized. There are two classes of documents—five about human–computer interaction (c1–c5) and four about mathematical graph theory (m1–m4).

(b)

X =

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

FIG. 7.2b. This dataset can be described by means of a term by document matrix in which each cell entry indicates the frequency with which a term occurs in a document.

7.2b. The linear decomposition is shown next (Fig. 7.3a), and the fact that its cross multiplication perfectly reconstructs the original is illustrated. Next we show a reduction to just two dimensions (Fig. 7.3b) that approximates the original matrix. This two dimensional approximation also allows us to give a geometrical representation of the dimensional structure, as shown in Fig. 7.4. The two-dimensional vector for each document and each word defines a point on a plane for each, as shown in the figure. The same space accommodates points representing both terms and documents. (To be technically precise, the distances between terms and between documents are correct in this figure, but those between terms and documents require a scaling operation. The approximation is close enough for illustration.)

The five human computer interaction documents are all in one part of the space, and the graph theory ones in another. A query can be represented in the same way as a point in the space, and one, "*human computer* interaction," is shown. Usually, we measure the similarity of a query to the documents by the cosine, or angle of its vector with respect to those of the documents. So in the figure, the query has a cone around it containing all points with cosine greater than .9. This region cleanly separates the titles to which the query is relevant from the ones to which it is not. A notable fact is that this includes

X = T * S * D′

T=

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

S=

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

D′=

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

FIG. 7.3a. The full dimensional Singular Value Decomposition of the matrix of Fig. 7.2b.

0.22	-0.11
0.20	-0.07
0.24	0.04
0.40	0.06
0.64	-0.17
0.27	0.11
0.27	0.11
0.30	-0.14
0.21	0.27
0.01	0.49
0.04	0.62
0.03	0.45

*

3.34	
	2.54

*

0.20	0.61	0.46	0.54	0.28
-0.06	0.17	-0.13	-0.23	0.11

= X^

	c1	c2	c3	c4
human	0.16	0.40	0.38	0.
interface	0.14	0.37	0.33	0.
computer	0.15	0.51	0.36	0.
user	0.26	0.84	0.61	0.
system	0.45	1.23	1.05	1.
response	0.16	0.58	0.38	0.
time	0.16	0.58	0.38	0.
EPS	0.22	0.55	0.51	0.
survey	0.10	0.53	0.23	0.
trees	-0.06	0.23	-0.14	-0.
graph	-0.06	0.34	-0.15	-0.
minors	-0.04	0.25	-0.10	-0.

FIG. 7.3b. The reduced two-dimer
7.2b.

relevant titles that contain none of
User response time and C3, *User in*
step has collapsed the meaning of
such a degree that documents that
contain other terms that originally
predicted to be likely to have incluc

Very roughly and anthropomorp
orthogonal dimensions to go on, h
in each cell. It does that by sayir
having so much of Factor one and
word has so much of Factor one ar
those two pieces of information (b
word X actually appeared 0.6 time

Comparing the rows for *huma*
two-dimensionally reconstructed m
were totally uncorrelated in the ori
the same document—they are qui
constructed approximation. Thus,

nents require a scaling operation. The
istration.)

ion documents are all in one part of the
another. A query can be represented in
nd one, "*human computer* interaction,"
milarity of a query to the documents by
respect to those of the documents. So in
nd it containing all points with cosine
eparates the titles to which the query is
not. A notable fact is that this includes

.11	-0.34	0.52	-0.06	-0.41
.28	0.50	-0.07	-0.01	-0.11
.11	-0.25	-0.30	0.06	0.49
.33	0.38	0.00	0.00	0.01
.16	-0.21	-0.17	0.03	0.27
.08	-0.17	0.28	-0.02	-0.05
.08	-0.17	0.28	-0.02	-0.05
.11	0.27	0.03	-0.02	-0.17
.54	0.08	-0.47	-0.04	-0.58
.59	-0.39	-0.29	0.25	-0.23
.07	0.11	0.16	-0.68	0.23
.30	0.28	0.34	0.68	0.18

50				
	1.31			
		0.85		
			0.56	
				0.36

.28	0.00	0.01	0.02	0.08
.11	0.19	0.44	0.62	0.53
.51	0.10	0.19	0.25	0.08
.15	0.02	0.02	0.01	-0.03
.33	0.39	0.35	0.15	-0.60
.03	-0.30	-0.21	0.00	0.36
.67	-0.34	-0.15	0.25	0.04
.06	0.45	-0.76	0.45	-0.07
.26	-0.62	0.02	0.52	-0.45

gular Value Decomposition of the matrix

0.22	-0.11
0.20	-0.07
0.24	0.04
0.40	0.06
0.64	-0.17
0.27	0.11
0.27	0.11
0.30	-0.14
0.21	0.27
0.01	0.49
0.04	0.62
0.03	0.45

*

3.34	
	2.54

*

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53

= X^

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

FIG. 7.3b. The reduced two-dimensional approximation to the matrix in Fig. 7.2b.

relevant titles that contain none of the words in the query, for example C5, *User response time* and C3, *User interface system.* The dimension reduction step has collapsed the meaning of words that appear in similar contexts to such a degree that documents that did not contain a particular term, but did contain other terms that originally occurred in similar contexts, are now predicted to be likely to have included that term.

Very roughly and anthropomorphically, SVD, with only values along two orthogonal dimensions to go on, has to guess what words actually appear in each cell. It does that by saying, "This document is best described as having so much of Factor one and so much of much of Factor two, and this word has so much of Factor one and so much of Factor two, and combining those two pieces of information (by vector arithmetic), my best guess is that word X actually appeared 0.6 times in document Y."

Comparing the rows for *human* and *user* in the original and in the two-dimensionally reconstructed matrices (Fig. 7.3) shows that although they were totally uncorrelated in the original—the two words never appeared in the same document—they are quite strongly correlated (r = .9) in the reconstructed approximation. Thus, SVD has done just what we wanted. It

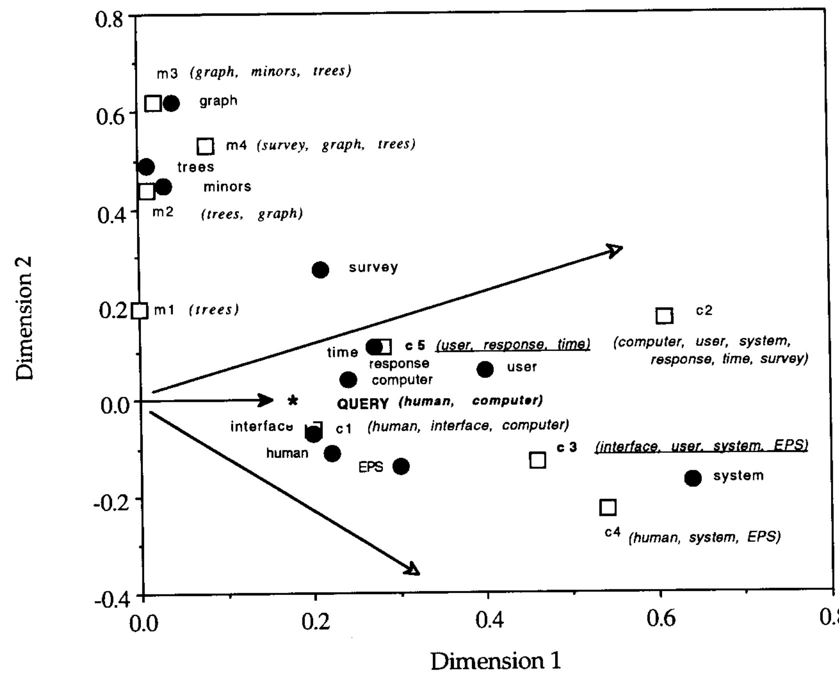


FIG. 7.4. A two-dimensional plot of 12 terms and nine documents from the sample of titles given in Fig. 7.2a and represented by the matrices of Fig. 7.3b. Terms are shown as filled circles. Documents are shown as open squares, their component terms indicated parenthetically. The query "*human computer* interaction" is represented as a pseudo-document. The cone represents the region within which points have a cosine of 0.9 or greater with the query. All documents about human–computer interaction (c1–c5) and none about graphs (m1–m4) are within this cone. In this reduced space, even documents c3 and c5, which share no terms with it, are near the query. (Axes are scaled for document–document or term–term comparisons.)

has filled in the documents with partial values for words that might well have been used in particular documents but weren't.

The shaded cell entries under m4 show this phenomenon in a slightly different way. The word *tree* did not appear in graph theory title m4. But because m4 did contain *graph* and *minor* the zero entry for *tree* has been replaced with 0.66, an estimate of how many times it occurs in titles containing *graph* and *minor*. By contrast, the value 1.00 for *survey*, which appeared once in m4, has been replaced by 0.42, reflecting the fact that it is unexpected in this context and should be counted as unimportant in matching a query. Notice that if we were to change the entry in any one cell of the original matrix, the values in the dimension reduced reconstruction would be changed everywhere.

When thus applied to informatio
Latent Semantic Indexing (LSI). LSI h
for. It relieves the user somewhat c
same words used by the author or in
has occurred in similar word contex
the query will stand a good chance c
ing and of rejecting documents of d

LSI's Information Retrieval Perfo

How well does all this work? In the
automatic retrieval of electronically s
significant improvement over prior m
problems. Its first tests were against
which representative queries have b
made more or less exhaustive search
which items are and are not releva
performance ranged from just equi
standard vector method with optimal
In a recent competition staged by t
Technology, LSI was compared wi
prototypes and commercial retrieval
sons among the many systems were s
amounts of preprocessing—things li
differences in stop lists, and the amou
before the final test runs. Nevertheless
to earlier ones. Compared to the star
was a 16% improvement (Dumais, 19

What does this mean? Approxima
mation retrieval is that when half of
found, less than half of the items yo
when half of the items you want ha
you found were wanted. Thus, altho
leaves much room for improvement

The LSI method has been appli
generally with results that pleased
objective measures were available, h
One interesting application to which
been applied is indexing across langu
this, LSI requires a training set of docu
or more languages. For each docume
the words from both languages is c
dimensions that are required for a sin
for each word in a common languag

When thus applied to information retrieval, we call the SVD technique Latent Semantic Indexing (LSI). LSI has just the property that we're looking for. It relieves the user somewhat of the need to know and produce the same words used by the author or indexer. If the user thinks of a word that has occurred in similar word contexts over the domain that was analyzed, the query will stand a good chance of matching documents of similar meaning and of rejecting documents of different meaning.

LSI's Information Retrieval Performance

How well does all this work? In the application for which it was designed, automatic retrieval of electronically stored document abstracts, it provides a significant improvement over prior methods but does not nearly solve all the problems. Its first tests were against standard collections of documents in which representative queries have been obtained and human judges have made more or less exhaustive searches of the whole database to determine which items are and are not relevant. In these standard collections LSI's performance ranged from just equivalent to the best prior method—the standard vector method with optimal term weighting—up to about 30% better. In a recent competition staged by the National Institute of Standards and Technology, LSI was compared with a large number of other research prototypes and commercial retrieval schemes. Direct quantitative comparisons among the many systems were somewhat muddied by the use of varying amounts of preprocessing—things like getting rid of typographical errors, differences in stop lists, and the amount of tweaking that systems were given before the final test runs. Nevertheless, the results appeared to be quite similar to earlier ones. Compared to the standard vector method *ceteris paribus* LSI was a 16% improvement (Dumais, 1994).

What does this mean? Approximately stated, the state of the art in information retrieval is that when half of the items you would want have been found, less than half of the items you have found were wanted. With LSI, when half of the items you want have been found, over 60% of the ones you found were wanted. Thus, although a significant step forward, LSI still leaves much room for improvement.

The LSI method has been applied in a variety of other applications, generally with results that pleased the designers and users and, where objective measures were available, has usually outperformed rival schemes. One interesting application to which no other fully automatic technique has been applied is indexing across languages (Landauer & Littman, 1990). To do this, LSI requires a training set of documents in which each is available in two or more languages. For each document a concatenated version containing all the words from both languages is constructed. Using the same number of dimensions that are required for a single language, the SVD result is a vector for each word in a common language-independent space. Given the mathematics of LSI, this would mean that any pair of words in the two languages that were used the same number of times in the same documents would have identical vectors, and that ones that are used in similar but not quite identical patterns across the documents will have similar but not quite identical vectors. Once the word vectors have been determined, they can be used for both new documents and new queries that are presented in only one of the training languages and will return appropriate documents in any language—once transformed into abstract numerical vectors, the system doesn't give a hoot which language either the query or document came from. For French and English paragraphs from the Canadian parliamentary proceedings, retrieval was as good for a query in one language finding documents in the other as it was for queries and documents in the same language. Almost as good results were obtained when going from Japanese ideographic *Kanji* characters to English words in a sample of technical abstracts.

LSI and Human Performance

The information retrieval results encouraged us to believe that LSI captures some of the underlying meaning structure of vocabulary when applied to large bodies of representative text. This presumption has been tested by predicting various aspects of the performance of human subjects dealing with textual materials.

Kintsch and his colleagues developed methods for representing text in a propositional language and have used them to analyze the coherence of discourse. They have shown that the comprehension of text depends heavily on its coherence—the continuity between the concepts expressed in one sentence or passage and the next. The Kintsch method requires difficult judgments by highly trained raters. This has limited research to very small samples of text and inhibited practical application to composition and instruction. Foltz, Kintsch, and Landauer (1993) tried applying LSI to the task. They started with a set of paragraphs about heart function that had been specifically constructed to have varying degrees of coherence, and for which comprehension measures had previously been obtained by testing students on their understanding of the texts. They obtained an LSI space by analyzing a collection of encyclopedia articles dealing with the heart. The LSI stand-in for coherence judgments was the cosine between each sentence and the following one. Fig. 7.5 shows the results. The LSI measure predicted comprehension scores extremely well, r = .93. For a control, we tried to predict comprehension using only the surface overlap of words, the first order correlation based on the proportion of word types in each sentence that were the same as those in the last. Technically this was realized as the cosine between successive sentences in the full-dimensional space, thus keeping everything constant except the dimension reduction step of the SVD analysis. This measure had almost no predictive value, $r = .18$.

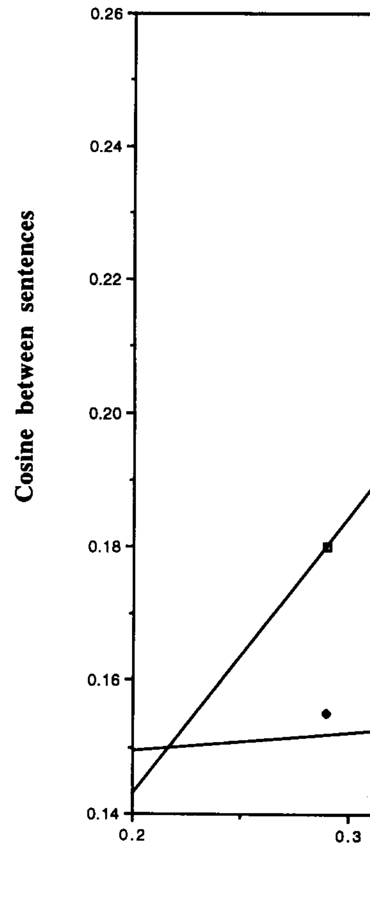


FIG. 7.5. Students' comprehensio
coherence. Coherence is measur
similarity (LSI) or by full dimensio
contained in successive sentences.

In a still-in-progress pilot study,
to use LSI to match students with
complexity for learning. Earlier wo
Kintsch, 1994) showed that people
is neither too hard, containing too
not yet familiar, nor too easy, cont
tested knowledge of cardiac func
characterized the typical student's
of correctly answered questions al
at how much students learned from

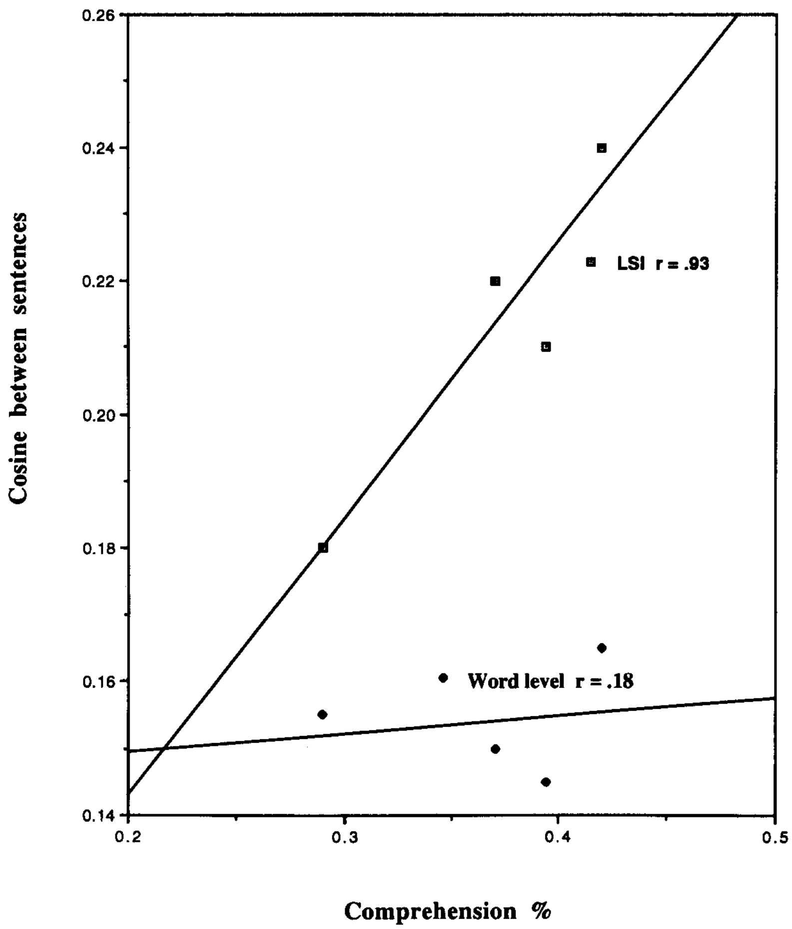


FIG. 7.5. Students' comprehension of text passages as related to their coherence. Coherence is measured either by reduced dimensional LSI similarity (LSI) or by full dimensional similarity (word level) of the words contained in successive sentences.

In a still-in-progress pilot study, Kintsch and Landauer have been trying to use LSI to match students with text at the optimal level of conceptual complexity for learning. Earlier work by Kintsch and his collaborators (see Kintsch, 1994) showed that people learn the most when the text on a topic is neither too hard, containing too many concepts with which a student is not yet familiar, nor too easy, containing too few. In a pilot study, Kintsch tested knowledge of cardiac function with a short-answer test. He then characterized the typical student's knowledge by the LSI vector for all text of correctly answered questions along with their answers. He then looked at how much students learned from reading text at varying levels of sophistication. He characterized the differentially sophisticated text by the centroid of the LSI vectors for the words it contained. When the vector point for students most closely matched that of the text, learning was greatest.

In yet another pilot study, Kintsch asked students to rate how familiar, memorable, and interesting the various paragraphs about heart function were. He found that the higher the cosine with the LSI measure of the typical student's knowledge, the more memorable, familiar, and interesting a paragraph appeared. These results are extremely preliminary and should be taken only as an additional indication that the LSI representation captures important aspects of meaning.

LSI and Synonym Tests

The initial purpose of LSI was to overcome the problem of synonymy in word usage for information retrieval. It has been our presumption and claim that the technique represents words of similar meaning in similar ways. When one compares words with similar vectors as derived from large collections, the claim is largely but not entirely fulfilled at an intuitive level. Many of the near neighbors of a word are indeed good synonyms, for example in the English–French cross-language indexing trial, the words *chambre* and *house* were quite close, as they should be in parliamentary usage. Most near neighbors, words with cosines over about .5, appear closely related in some manner. In a scaling of an encyclopedia, *surgeon*, *physician*, *patient*, and *bedside* are all close to one another. But the relationship between some close neighbors in LSI space can occasionally be quite mysterious (e.g., *verbally* and *sadomasochism* with a cosine of .8). It's impossible to say exactly why, but it's plausible that some words that have more than one meaning receive a sort of average value that signifies nothing, and that many words are sampled too thinly to get well placed. It's also possible, of course, that the "bag of words" method, which ignores all syntactical and logical entailments, sometimes misses meaning or gets it scrambled.

We were interested to see how well, compared to people, LSI captures synonymy. To do so, we measured LSI's knowledge of synonyms on a standardized test. The test was taken from the ETS Test of English as a Foreign Language (TOEFL). (It is worth noting that ETS does not use general synonym tests for ordinary verbal ability assessment because they are too easy for college students.) To make these comparisons, we first trained LSI by running the analysis on a large corpus of representative English. In various studies, we have used both collections of newspaper text from the Associated Press news wire and *Grolier's Academic American Encyclopedia*, a work intended for students. In the most successful study, we performed an SVD on segments consisting of the first 2,000 characters or less (on average, 152 words) of each of 30,473 articles in the encyclopedia. This resulted in a vector for each of 60,768 words.

The TOEFL vocabulary test consists
a single word, and there are four alte
among which the test taker is to ch
gave a prediction of the best alternati
the stem, in 74 of the 80 test items.
never met either the stem word and/
guess with probability .25. Scored th
Average test takers, students applyin
States from non-English-speaking cou
TOEFL we used. Thus, having "rea
encyclopedia, LSI did as well as the

To dot some *i*s, in this study we a
from 200 to 372 (the number at whic
matically terminated; see Berry, 1992
weakly nonmonotonic trends we are
tests: 51.5 correct with 300 and 325 di
fewer (225) or more (372) dimension

We also compared the pattern of er
question we computed a product-mom
cosine of the stem and each alternative
alternative in a large sample of student
items was 0.70. Excluding the correct
.44, showing that LSI confusions are
those of students. When LSI choose
correctly, it sometimes appears to be b
tual associations and less to contrasti
prefers *nurse* (cos = .47) to *doctor* (co

In an important control experiment
by the degree of surface co-occurren
sages. We did this by applying a stan
each word is treated as independent;
reduction. Choosing the best alternat
29.5 (37%) correct answers. This demo
reduction technique captures more th

[4]From an AI or linguistics perspective, one m
word types is a straw control, in that derivatio
be counted as equivalent. However, the poin
simulate how the very knowledge such relation
equivalence relations about form variants in the
by their similarity in the dimension-reduced S
similarities and morphemic combinatorics pla
understanding of words, but we have not yet fo
Preliminary attempts in which we added comp
in the LSI matrix produced only degradation in

ntially sophisticated text by the centroid
contained. When the vector point for
of the text, learning was greatest.
h asked students to rate how familiar,
us paragraphs about heart function were.
e with the LSI measure of the typical
orable, familiar, and interesting a para-
remely preliminary and should be taken
ne LSI representation captures important

vercome the problem of synonymy in
It has been our presumption and claim
s of similar meaning in similar ways.
nilar vectors as derived from large col-
t entirely fulfilled at an intuitive level.
word are indeed good synonyms, for
ss-language indexing trial, the words
se, as they should be in parliamentary
ith cosines over about .5, appear closely
of an encyclopedia, *surgeon, physician,*
e another. But the relationship between
can occasionally be quite mysterious
with a cosine of .8). It's impossible to
t some words that have more than one
ue that signifies nothing, and that many
well placed. It's also possible, of course,
hich ignores all syntactical and logical
ing or gets it scrambled.
well, compared to people, LSI captures
d LSI's knowledge of synonyms on a
en from the ETS Test of English as a
th noting that ETS does not use general
bility assessment because they are too
these comparisons, we first trained LSI
pus of representative English. In various
s of newspaper text from the Associated
lemic American Encyclopedia, a work
uccessful study, we performed an SVD
000 characters or less (on average, 152
n the encyclopedia. This resulted in a

The TOEFL vocabulary test consists of items in which the stem is usually a single word, and there are four alternatives, usually single word answers, among which the test taker is to choose the one most like the stem. LSI gave a prediction of the best alternative, the one with the highest cosine to the stem, in 74 of the 80 test items. For the remaining six, where LSI had never met either the stem word and/or the correct alternative, we made it guess with probability .25. Scored this way, LSI got 51.5 (64.5%) correct. Average test takers, students applying for college entrance in the United States from non-English-speaking countries, got 51.6 correct on the form of TOEFL we used. Thus, having "read" 4.6 million words from a general encyclopedia, LSI did as well as the average foreign student.

To dot some *i*s, in this study we also varied the number of dimensions from 200 to 372 (the number at which the decomposition algorithm automatically terminated; see Berry, 1992). The TOEFL test results showed the weakly nonmonotonic trends we are accustomed to in information retrieval tests: 51.5 correct with 300 and 325 dimensions, and 47.5 correct with either fewer (225) or more (372) dimensions.

We also compared the pattern of errors of LSI to that of students. For each question we computed a product-moment correlation coefficient between the cosine of the stem and each alternative with the proportion of guesses for each alternative in a large sample of students. The average correlation across the 80 items was 0.70. Excluding the correct alternative, the average correlation was .44, showing that LSI confusions are somewhat like and somewhat unlike those of students. When LSI chooses wrongly and most students choose correctly, it sometimes appears to be because LSI is more sensitive to contextual associations and less to contrastive features. For example, LSI slightly prefers *nurse* (cos = .47) to *doctor* (cos = .41) as an associate to *physician.*

In an important control experiment, we chose the correct answers simply by the degree of surface co-occurrence of words in the encyclopedia passages. We did this by applying a standard vector retrieval method in which each word is treated as independent; that is, there is there is no dimension reduction. Choosing the best alternative by the highest cosine yielded just 29.5 (37%) correct answers. This demonstrates once again that the dimension reduction technique captures more than mere co-occurrence.[4] More impor-

[4]From an AI or linguistics perspective, one might object that first-order correlations between word types is a straw control, in that derivational and inflectional variants of words ought to be counted as equivalent. However, the point of the present endeavor is to understand or simulate how the very knowledge such relations imply is acquired by experience. LSI acquires equivalence relations about form variants in the same way it acquires those based on synonymy, by their similarity in the dimension-reduced SVD space. It seems likely that morphological similarities and morphemic combinatorics play an additional role in human learning and understanding of words, but we have not yet found a successful way to model such processes. Preliminary attempts in which we added component letter *n*-grams to the words represented in the LSI matrix produced only degradation in the TOEFL test results.

tant for our next argument, it implies that indirect associations or structural relations induced by analysis of the whole corpus are involved in LSI's success with individual words. Thus, correct representation of any one word may depend on the correct representation of many, perhaps all other words.

LSI and the Vocabulary Learning Paradox

LSI is doing a pretty good job of mimicking human performance. We like to say, only partly tongue in cheek, that it is doing real artificial intelligence. It has been learning lexical semantics entirely automatically, entirely artificially. No one has plugged in semantic information from their own heads, as is done in all other natural language understanding systems, and no preexisting humanly constructed dictionary or thesaurus is involved. The system has only a mathematical machine that it uses to run over text and extract knowledge on its own. The test of semantic knowledge that we have given it is one that is central to tests of human intelligence, vocabulary being the single measure that best correlates with overall verbal intelligence and scholastic achievement. Knowledge of words and the concepts for which they stand is at once the major foundation of human intelligence and its crowning achievement.

How does LSI's learning rate—the number of words that it "knows" as a function of how much it has "read"—compare with humans' learning rates? To know whether LSI has actually matched humans in acquisition rate, words learned per word read, we would need to know how many words of text in English the average TOEFL taker has met. This we do not know. Reading at 165 words per minute—the average for U.S. fifth graders—it would take 468 hours to read 4.6 million words. Very informal questioning of a few foreign students has suggested that they have read something like that amount of English text. An average schoolchild will have read 4.6 million words by around seventh grade. Does an average seventh grader's vocabulary equal that of TOEFL takers? We don't know. The most satisfactory answer would be obtained by testing grade-schoolers on the same items, which we have not done.

Here's another approach to the LSI—human comparison. One way a test-taker could get a word right on the TOEFL test would be by knowing the meaning of both the stem word and the correct alternative. By this model the proportion of words known is the square root of the probability correct; for our data 72%, corrected for guessing. If the TOEFL words were a random sample of words in the encyclopedia, this would mean that LSI had learned about 72% of the word types it had read, or roughly 44,000. This would give a learning rate of about one word learned per hundred total tokens read. This is about 1.6 times what children achieve naturally.

Here's yet another approach to the LSI—human comparison. If we consider only those items where LSI had met both stem and correct alternative at least

s that indirect associations or structural
e whole corpus are involved in LSI's
correct representation of any one word
ation of many, perhaps all other words.

Paradox

nicking human performance. We like to
t it is doing real artificial intelligence. It
ntirely automatically, entirely artificially.
mation from their own heads, as is done
rstanding systems, and no preexisting
esaurus is involved. The system has only
to run over text and extract knowledge
wledge that we have given it is one that
ce, vocabulary being the single measure
bal intelligence and scholastic achieve-
concepts for which they stand is at once
ligence and its crowning achievement.
e number of words that it "knows" as a
—compare with humans' learning rates?
matched humans in acquisition rate,
would need to know how many words
L taker has met. This we do not know.
—the average for U.S. fifth graders—it
illion words. Very informal questioning
sted that they have read something like
ge schoolchild will have read 4.6 million
es an average seventh grader's vocabu-
We don't know. The most satisfactory
ng grade-schoolers on the same items,

LSI—human comparison. One way a
the TOEFL test would be by knowing
nd the correct alternative. By this model
e square root of the probability correct;
sing. If the TOEFL words were a random
a, this would mean that LSI had learned
d read, or roughly 44,000. This would
word learned per hundred total tokens
children achieve naturally.
e LSI—human comparison. If we consider
both stem and correct alternative at least

once, the proportion of items correct, adjusted for guessing, was .57. By fitting Bower's (1961) one-trial learning model (which for this purpose is equivalent to many others) to the data, we estimated a learning rate, the probability of going from wrong to right on one exposure. The fitting procedure used a hill-climbing algorithm to find a learning rate parameter that, when applied to the number of occurrences of each stem and correct alternative pair for TOEFL items, gave an average proportion correct of .57. The estimated learning rate was .049, almost equal to Nagy's rough estimate of 1 in 20 for children learning from natural context in the lab (Nagy & Herman, 1987).

Despite the obviously wide uncertainty of these estimates, it seems clear enough that LSI acquires word knowledge at a clip that approaches the comparable human achievement. More important, LSI is using a mechanism for the acquisition of word knowledge that is not represented in any of our theories of memory. Notice that LSI was deprived of any use of syntax, logic, morphology, perceptual grounding, or real-world knowledge and pragmatics. Surely some of these must be of some help to humans. Suppose we were able to add these sources of information to what LSI could do. Would the high rates of learning of which humans are capable remain mysterious? We think not.

What has LSI machinery added that was undreamed of in our theories and our methods of vocabulary instruction? LSI does indirect learning. That is, it can improve its knowledge of a word in its absence and in the absence of any word that looks or sounds like it (LSI has no knowledge of what words look or sound like), the usual mechanism of positive transfer in learning theories. A vector assigned to any particular word is an average of the vectors of all the words in all the passages in which it occurs. The vectors of those words in turn are averages of all of the words with which they have kept company. The "meaning" of word X is greatly influenced by things that happen in paragraphs in which X does not appear. Indeed the occasions on which word X is actually found are a minor part of the experience by which its meaning is defined. If we imagined an LSI-like process to be working sequentially over a lifetime, experiences with other words before and after it is met, when it is not present, will have great bearing on what it "means." In laboratory attempts to teach children new words by embedding them in context, all this indirect learning is absent, or at least unmeasured. Put differently, when a child is presented with a paragraph containing one new word there may be much more learning than is measured by testing that word. There is learning about all the words in the paragraph and about their entailments with all words in the English language, or at least so the LSI model would have it.

We tested this property of the model more directly. Suppose we continue to give LSI all the documents that contain words from a particular test item, but reduce the number of other documents it sees. If there are indirect effects, the chances of a correct answer should be smaller. We reran the LSI analysis, excluding 17,394 encyclopedia articles that contained no words—either stems or answer alternatives—from 20 selected items, and had it take the test again.[5] Twelve of the 20 were correct with all the documents, only 4 with the reduced document set, ($p = .01$ by exact test). In other words, depriving the LSI analysis of part of the data about words not on tests of the items seriously diminished its measured knowledge of test words.[6]

This result demonstrates a kind of generalization or transfer that arises entirely from similarity relations derived from co-occurrence experience, with no contribution from or grounding in preexisting perceptual or categorical primitives, and no exogenous reinforcement of the correctness of the inferred relations among the atomic units.

The mathematical machinery of SVD is capable of making inferences based on an underlying structure in the use of words that increases learning power. Is there a similar machine in the human mind? If so, one would expect it to apply not only to the acquisition of word meanings but to the acquisition and representation of knowledge in all domains. Word meanings are particularly central, because much of what we know about any topic is contained or reflected in what we know about its vocabulary. LSI's knowledge of words is limited to analysis and prediction of the commonality of the contexts in which they occur. This seems a far cry from what we have usually imagined to be the structure of most knowledge. But is it? How much of what the average student—or average professor—knows about history, geography or botany lies in just in these same kinds of direct and indirect associations? How much of useful knowledge-based performance is knowing the right word to think or say in the right verbal context? The answer is not obvious.

As a potential theory of memory, LSI also has some intuitively intriguing qualitative properties. For one example, why do parents mix up their children's names, even when in the presence of just one child and away from home, so that the classical stimulus overload explanation falters? LSI says the reduced-dimensional representation of two siblings' names are likely to be almost identical. For another example, what, exactly, do we mean when we say that no two words have exactly the same meaning, that a word never has the same meaning on different occasions or for different people? LSI offers the hope of saying more exactly what that means.

[5]The selected items were those that changed scores from right to wrong or wrong to right between an analysis based on a random subset of 10,000 documents, for which the overall number correct was 39.5, and the full sample of 30,473 documents ($p = .02$ by exact test).

[6]The average document that did contain those TOEFL words had just 5 such words out of 152 total tokens. Thus the remaining 13,079 documents on which the reduced context TOEFL test was based still contain a great deal of information about words other than the TOEFL terms. The reduced set produced vectors for approximately 39,320 non-TOEFL word types, compared to 60,400 for the original analysis.

There are several appealing next
LSI as a theory of knowledge and me
in the laboratory. One might const
associative matrices (much like the o
students appropriately to the inforn
their recollections fill in missing cell
filled cells in the ways predicted by
to extend the studies done here from
edge in various fields, to see wheth
textbook, LSI can do well on multip
geography. Still another line would
textual data to mimic other phenom
association norms and analogy judg

At this juncture, we are not yet con
new theory of knowledge or memory
of human memory. Nonetheless, we
account for some of the mysteriou
represent vast quantities of informati
prior theories, which do not contain
Thus, the question is raised for memo
learning mechanism similar to the c
candidate for biological memory? Su
knowledge acquisition relies to a su
An important consequence for educ
to teach individual unknown words
reading of rich and varied content w

We have come full circle. We hav
a case in which research into a pract
neering of an aid for external memo
memory works and has suggested ne
that, in turn, may imply strategies fo

REFE

Anderson, R. C., & Freebody, P. (1983). Re
knowledge. In B. Hutson (Ed.), *Advances i*
(pp. 231–256). Greenwich, CT: JAI Press.

Anderson, R. C., Wilson, P. T., & Fielding, L.
spend their time outside of school. *Readin*

Berry, M. W. (1992). Large scale singular valu
computer Applications, 6, 13–49.

Bower, G. H. (1961). Application of a model
255–280.

Carroll, J. B., Davies, P., & Richman, B. (1971
Boston: Houghton-Mifflin.

r should be smaller. We reran the LSI
ia articles that contained no words—
om 20 selected items, and had it take
correct with all the documents, only
= .01 by exact test). In other words,
he data about words not on tests of
ısured knowledge of test words.[6]

generalization or transfer that arises
ved from co-occurrence experience,
ng in preexisting perceptual or cate-
reinforcement of the correctness of
ic units.

VD is capable of making inferences
e use of words that increases learning
the human mind? If so, one would
uisition of word meanings but to the
vledge in all domains. Word meanings
of what we know about any topic is
ow about its vocabulary. LSI's knowl-
nd prediction of the commonality of
s seems a far cry from what we have
of most knowledge. But is it? How
-or average professor—knows about
ıst in these same kinds of direct and
ıseful knowledge-based performance
r say in the right verbal context? The

SI also has some intuitively intriguing
le, why do parents mix up their chil-
ence of just one child and away from
overload explanation falters? LSI says
on of two siblings' names are likely to
ple, what, exactly, do we mean when
ctly the same meaning, that a word
ent occasions or for different people?
xactly what that means.

d scores from right to wrong or wrong to right
et of 10,000 documents, for which the overall
of 30,473 documents (p = .02 by exact test).
ose TOEFL words had just 5 such words out of
cuments on which the reduced context TOEFL
formation about words other than the TOEFL
approximately 39,320 non-TOEFL word types,

There are several appealing next steps in pursuing the implications of LSI as a theory of knowledge and memory. First, would be to test it rigorously in the laboratory. One might construct artificial vocabularies or artificial associative matrices (much like the ones for words and documents), expose students appropriately to the information in the matrix, and see whether their recollections fill in missing cells and alter the probability of recall of filled cells in the ways predicted by SVD analysis. Another tack would be to extend the studies done here from word knowledge to substantive knowledge in various fields, to see whether having "read" an encyclopedia or a textbook, LSI can do well on multiple-choice questions about history and geography. Still another line would be to attempt to use the theory and textual data to mimic other phenomena of human word memory, such as association norms and analogy judgments.

At this juncture, we are not yet convinced that we have hit on an important new theory of knowledge or memory, or even an important new mechanism of human memory. Nonetheless, we have identified a mechanism that can account for some of the mysterious power of the mind to acquire and represent vast quantities of information, and we have reason to believe that prior theories, which do not contain the same mechanism, are inadequate. Thus, the question is raised for memory theory: Should an indirect inferential learning mechanism similar to the one employed by LSI be considered a candidate for biological memory? Suppose the answer is yes, that natural knowledge acquisition relies to a substantial degree on similar processes. An important consequence for education would follow: It is not sufficient to teach individual unknown words or concepts in isolation; voluminous reading of rich and varied content would appear more promising.

We have come full circle. We have taken the first steps in what may be a case in which research into a practical problem—one involving the engineering of an aid for external memories—has raised issues of how internal memory works and has suggested new forms of theory for natural memory that, in turn, may imply strategies for practical memory problems.

REFERENCES

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research: A research annual* (pp. 231–256). Greenwich, CT: JAI Press.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23*, 285–303.

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Applications, 6*, 13–49.

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrica, 26*, 255–280.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston: Houghton-Mifflin.

Carver, R. P. (1990). *Reading rate: A review of research and theory.* San Diego: Academic Press.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society For Information Science, 41,* 391–407.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), *National Institute of Standards and Technology Text Retrieval Conference,* NIST special publication.

Durkin, D. (1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly, 14,* 481–253.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1993). An analysis of textual coherence using Latent Semantic Indexing. *Society for Text and Discourse.*

Furnas, G. W. (1985). Experience with an adaptive indexing scheme. In *Proceedings of CHI'85* (pp. 16–23). New York: ACM.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983a). Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal, 62,* 1753–1804.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30,* 964–971.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21,* 767–787.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49,* 294–303.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change.* Chicago: University of Chicago Press.

Landauer, T. K. (1986). How much do people remember: Some estimates of the quantity of learned information in long-term memory. *Cognitive Science, 10,* 477–493.

Landauer, T. K., & Littman, M. L. (1990) Fully automatic cross-language document retrieval using latent semantic indexing. In *Conference on Electronic Text Research.* Waterloo, Canada.

Mokyr, J. (1990). *The lever of riches: Technological creativity and economic progress.* New York: Oxford University Press.

Nagy, W., & Anderson, R. (1984). The number of words in printed school English. *Reading Research Quarterly, 19,* 304–330.

Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. C. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19–35). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly, 20,* 223–253.

Petroski, H. (1982). *To engineer is human.* New York: Random House.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: Morrow.

Pressley, M., Ross, K. A., Levin, J. R., & Ghatala, E. S. (1984). The role of strategy utility knowledge in children's strategy decision making. *Journal of Experimental Child Psychology, 38,* 491–504.

Smith, M. (1941). Measurement of the size of general English vocabulary through the elementary grades and high school. *Genetic Psychology Monographs, 24,* 311–345.

Taylor, B. M., Frye, B. J., & Maruyama, G. M. (1990). Time spent reading and reading growth. *American Educational Research Journal, 27,* 351–362.

Walker, D. E., & Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In R. Grisham & R. Kittredge (Eds.), *Analyzing languages in restricted domains: Sublanguage description and processing.* Hillsdale, NJ: Lawrence Erlbaum Associates.

What Are the
of Individu
in Mem

Patri
Qi
University of Manch

People differ markedly in memor
Nevertheless, there are no satisfact
the efficiency of functional process
cognitive psychologists have been
differences associated with exten
Krampe, & Tesch-Rohmer, 1993)
niques (Ericcsen & Chase, 1982; E
for how the efficiency of the funct
vary between individuals polarize
without any cross-reference. One p
nitive skills, including memory, loa
ability such as Spearman's (1927) "
1984). Interest in this somewhat
gestion that "*g*" can be reified i
parameter of the functional cognit
formation processing speed (Ande
& Deary, 1982; Eysenck, 1986; Hu
telbeck, 1982; Smith & Stanley, 19
1984). A more recent is "working n
1990; Kyllonen & Crystal, 1990).
correlations between performance
(IQ Test Scores, IQTS) and perfor
ratory tasks such as Choice Reactio