An Introduction to Latent Semantic Analysis

Thomas K Landauer
Department of Psychology
University of Colorado at Boulder,

Peter W. Foltz
Department of Psychology
New Mexico State University

Darrell Laham
Department of Psychology
University of Colorado at Boulder,

Landauer, T. K., Foltz, P. W., & Laham, D. (1998).
Introduction to Latent Semantic Analysis.
Discourse Processes, **25**, 259-284.

Abstract

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer and Dumais, 1997). The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways. For example, its scores overlap those of humans on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word–word and passage–word lexical priming data; and, as reported in 3 following articles in this issue, it accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay.

An Introduction to Latent Semantic Analysis

Research reported in the three articles that follow—Foltz, Kintsch & Landauer (1998/this issue), Rehder, et al. (1998/this issue), and Wolfe, et al. (1998/this issue)—exploits a new theory of knowledge induction and representation (Landauer and Dumais, 1996, 1997) that provides a method for determining the similarity of meaning of words and passages by analysis of large text corpora. After processing a large sample of machine-readable language, Latent Semantic Analysis (LSA) represents the words used in it, and any set of these words—such as a sentence, paragraph, or essay—either taken from the original corpus or new, as points in a very high (e.g. 50-1,500) dimensional "semantic space". LSA is closely related to neural net models, but is based on singular value decomposition, a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people.

Word and passage meaning representations derived by LSA have been found capable of simulating a variety of human cognitive phenomena, ranging from developmental acquisition of recognition vocabulary to word-categorization, sentence-word semantic priming, discourse comprehension, and judgments of essay quality. Several of these simulation results will be summarized briefly below, and additional applications will be reported in detail in following articles by Peter Foltz, Walter Kintsch, Thomas Landauer, and their colleagues. We will explain here what LSA is and describe what it does.

LSA can be construed in two ways: (1) simply as a practical expedient for obtaining approximate estimates of the contextual usage substitutability of words in larger text segments, and of the kinds of—as yet incompletely specified— meaning similarities among words and text segments that such relations may reflect, or (2) as a model of the computational processes and representations underlying substantial portions of the acquisition and utilization of knowledge. We next sketch both views.

As a practical method for the characterization of word meaning, we know that LSA produces measures of word-word, word-passage and passage-passage relations that are well correlated with several human cognitive phenomena involving association or semantic similarity. Empirical evidence of this will be reviewed shortly. The correlations demonstrate close resemblance between what LSA extracts and the way peoples' representations of meaning reflect what they have read and heard, as well as the way human representation of meaning is reflected in the word choice of writers. As one practical consequence of this correspondence, LSA allows us to closely approximate human judgments of meaning similarity between words and to objectively predict the consequences of overall word-based similarity between passages, estimates of which often figure prominently in research on discourse processing.

It is important to note from the start that the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations (thus the phrase "Latent Semantic"), and as a consequence are often much better predictors of human meaning-based judgments and performance than are the surface level contingencies that have long been rejected (or, as Burgess and Lund, 1996 and this volume, show, unfairly maligned) by linguists as the basis of language phenomena.

LSA, as currently practiced, induces its representations of the meaning of words and passages from analysis of text alone. None of its knowledge comes directly from perceptual information about the physical world, from instinct, or from experiential intercourse with bodily functions, feelings and intentions. Thus its representation of reality is bound to be somewhat sterile and bloodless. However, it does take in descriptions and verbal outcomes of all these juicy processes, and so far as writers have put such things into words, or that their words have reflected such matters unintentionally, LSA has at least potential access to knowledge about them. The representations of passages that LSA forms can be interpreted as abstractions of "episodes", sometimes of episodes of purely verbal content such as philosophical arguments, and sometimes episodes from real or imagined life coded into verbal descriptions. Its representation of words, in turn, is intertwined with and mutually interdependent with its knowledge of episodes. Thus while LSA's potential knowledge is surely imperfect, we believe it can offer a close enough approximation to people's knowledge to underwrite theories and tests of theories of cognition. (One might consider LSA's maximal knowledge of the world to be analogous to a well-read nun's knowledge of sex, a level of knowledge often deemed a sufficient basis for advising the young.)

However, LSA as currently practiced has some additional limitations. It makes no use of word order, thus of syntactic relations or logic, or of morphology. Remarkably, it manages to extract correct reflections of passage and word meanings quite well without these aids, but it must still be suspected of resulting incompleteness or likely error on some occasions.

LSA differs from some statistical approaches discussed in other articles in this issue and elsewhere in two significant respects. First, the input data "associations" from which LSA induces representations are between unitary expressions of meaning—words and complete meaningful utterances in which they occur—rather than between successive words. That is, LSA uses as its initial data not just the summed contiguous pairwise (or tuple-wise) co-occurrences of words but the detailed patterns of occurrences of very many words over very large numbers of local meaning-bearing contexts, such as sentences or paragraphs, treated as unitary wholes. Thus it skips over how the order of words produces the meaning of a sentence to capture only how differences in word choice and differences in passage meanings are related.

Another way to think of this is that LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains. LSA's ability to simultaneously—conjointly—derive representations of these two interrelated kinds of meaning depends on an aspect of its mathematical machinery that is its second important property. LSA assumes that the choice of dimensionality in which all of the local word-context relations are simultaneously represented can be of great importance, and that reducing the dimensionality (the number parameters by which a word or passage is described) of the observed data from the number of initial contexts to a much smaller—but still large—number will often produce much better approximations to human cognitive relations. It is this dimensionality reduction step, the combining of surface information into a deeper abstraction, that captures the mutual implications of words and passages. Thus, an important component of applying the technique is finding the optimal dimensionality for the final representation. A possible interpretation of this step, in terms more familiar to researchers in psycholinguistics, is that the resulting dimensions of description are analogous to the semantic features often postulated as the basis of word meaning, although establishing concrete relations to mentalisticly interpretable features poses daunting technical and conceptual problems and has not yet been much attempted.

Finally, LSA, unlike many other methods, employs a preprocessing step in which the overall distribution of a word over its usage contexts, independent of its correlations with other words, is first taken into account; pragmatically, this step improves LSA’s results considerably.

However, as mentioned previously, there is another, quite different way to think about LSA. Landauer and Dumais (1997) have proposed that LSA constitutes a fundamental computational theory of the acquisition and representation of knowledge. They maintain that its underlying mechanism can account for a long-standing and important mystery, the inductive property of learning by which people acquire much more knowledge than appears to be available in experience, the infamous problem of the "insufficiency of evidence" or "poverty of the stimulus." The LSA mechanism that solves the problem consists simply of accommodating a very large number of local co-occurrence relations (between the right kinds of observational units) simultaneously in a space of the right dimensionality. Hypothetically, the optimal space for the reconstruction has the same dimensionality as the source that generates discourse, that is, the human speaker or writer's semantic space. Naturally observed surface co-occurrences between words and contexts have as many defining dimensions as there are words or contexts. To approximate a source space with fewer dimensions, the analyst, either human or LSA, must extract information about how objects can be well defined by a smaller set of common dimensions. This can best be accomplished by an analysis that accommodates all of the pairwise observational data in a space of the same lower dimensionality as the source. LSA does this by a matrix decomposition performed by a computer algorithm, an analysis that captures much indirect information contained in the myriad constraints, structural relations and mutual entailments latent in the local observations available to experience.

The principal support for these claims has come from using LSA to derive measures of the similarity of meaning of words from text. The results have shown that: (1) the meaning similarities so derived closely match those of humans, (2) LSA's rate of acquisition of such knowledge from text approximates that of humans, and (3) these accomplishments depend strongly on the dimensionality of the representation. In this and other ways, LSA performs a powerful and, by the human-comparison standard, correct induction of knowledge. Using representations so derived, it simulates a variety of other cognitive phenomena that depend on word and passage meaning.

The case for or against LSA's psychological reality is certainly still open. However, especially in view of the success to date of LSA and related models, it can not be settled by theoretical presuppositions about the nature of mental processes (such as the presumption, popular in some quarters, that the statistics of experience are an insufficient source of knowledge.) Thus, we propose to researchers in discourse processing not only that they use LSA to expedite their investigations, but that they join in the project of testing, developing and exploring its fundamental theoretical implications and limits.

What is LSA?

LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

The first step is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. Next, the cell entries are subjected to a preliminary transformation, whose details we will describe later, in which each cell frequency is weighted by a function that expresses both the word's importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general.

Next, LSA applies singular value decomposition (SVD) to the matrix. 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. One component matrix describes the original row entities as vectors of derived orthogonal factor values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-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. One can reduce the dimensionality of the solution simply by deleting coefficients in the diagonal matrix, ordinarily starting with the smallest. (In practice, for computational reasons, for very large corpora only a limited number of dimensions—currently a few thousand— can be constructed.)

Here is a small example that gives the flavor of the analysis and demonstrates what the technique accomplishes. This example uses as text passages the titles of nine technical memoranda, five about human computer interaction (HCI), and four about mathematical graph theory, topics that are conceptually rather disjoint. Thus 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 extracted terms italicized, and the corresponding word-by-document matrix is shown in **Figure 1**.[1] We will discuss the highlighted parts of the tables in due course.

The linear decomposition is shown next (**Figure 2**); except for rounding errors, its multiplication perfectly reconstructs the original as illustrated.

Next we show a reconstruction based on just two dimensions (**Figure 3**) that approximates the original matrix. This uses vector elements only from the first two, shaded, columns of the three matrices shown in the previous figure (which is equivalent to setting all but the highest two values in S to zero).

Each value in this new representation has been computed as a linear combination of values on the two retained dimensions, which in turn were computed as linear combinations of the original cell values. Note, therefore, that if we were to change the entry in any one cell of the original, the values in the reconstruction with reduced dimensions

---

[1] This example has been used in several previous publications (e.g. Deerwester et al., 1990; Landauer & Dumais, in press).

might be changed everywhere; this is the mathematical sense in which LSA performs inference or induction.

Example of text data: Titles of Some 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* management *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*

$\{X\} =$

	c1	**c2**	**c3**	**c4**	**c5**	**m1**	**m2**	**m3**	**m4**
human	1	0	0	1	0	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

r (human.user) = -.38

r (human.minors) = -.29

Figure 1. A word by context matrix, X, formed from the titles of five articles about human-computer interaction and four about graph theory. Cell entries are the number of times that a word (rows) appeared in a title (columns) for words that appeared in at least two titles.

The dimension reduction step has collapsed the component matrices in such a way that words that occurred in some contexts now appear with greater or lesser estimated frequency, and some that did not appear originally now do appear, at least fractionally.

$$\{X\} = \{W\}\{S\}\{P\}'$$

$$\{W\} =$$

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

$$\{P\} =$$

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

Figure 2. Complete SVD of matrix in Figure 1.

$$\left\{\hat{X}\right\} =$$

	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

r (human.user) = .94

r (human.minors) = -.83

Figure 3. Two dimensional reconstruction of original matrix shown in Fig. 1 based on shaded columns and rows from SVD as shown in Fig. 2. Comparing shaded and boxed rows and cells of Figs. 1 and 3 illustrates how LSA induces similarity relations by changing estimated entries up or down to accommodate mutual constraints in the data.

Look at the two shaded cells for survey and trees in column m4. The word tree did not appear in this graph theory title. But because m4 did contain graph and minors, the zero entry for tree has been replaced with 0.66, which can be viewed as an estimate of how many times it would occur in each of an infinite sample of titles containing graph and minors. 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 characterizing the passage. Very roughly and anthropomorphically, in constructing the reduced dimensional representation, SVD, with only values along two orthogonal dimensions to go on, has to estimate what words actually appear in each context by using only the information it has extracted. It does that by saying the following:

> This text segment is best described as having so much of abstract concept one and so much of abstract concept two, and this word has so much of concept one and so much of concept two, and combining those two pieces of information (by vector arithmetic), my best guess is that word X actually appeared 0.6 times in context Y.

Now let us consider what such changes may do to the imputed relations between words or between multi-word textual passages. For two examples of word-word relations, compare the shaded and/or boxed rows for the words human, user and minors (in this context, minor is a technical term from graph theory) in the original and in the two-dimensionally reconstructed matrices (**Figures 1 and 3**). In the original, human never appears in the same passage with either user or minors—they have no co-occurrences, contiguities or "associations" as often construed. The correlations (using Spearman r to facilitate familiar interpretation) are -.38 between human and user, and a slightly higher -.29 between human and minors. However, in the reconstructed two-dimensional approximation, because of their indirect relations, both have been greatly altered: the human-user correlation has gone up to .94, the human-minors correlation down to -.83. Thus, because the terms human and user occur in contexts of similar meaning—even though never in the same passage—the reduced dimension solution represents them as more similar, while the opposite is true of human and minors.

To examine what the dimension reduction has done to relations between titles, we computed the intercorrelations between each title and all the others, first based on the raw co-occurrence data, then on the corresponding vectors representing titles in the two-dimensional reconstruction; see **Figure 4**.

In the raw co-occurrence data, correlations among the 5 human-computer interaction titles were generally low, even though all the papers were ostensibly about quite similar topics; half the rs were zero, three were negative, two were moderately positive, and the average was only .02. The correlations among the four graph theory papers were mixed, with a moderate mean r of 0.44. Correlations between the HCI and graph theory papers averaged only a modest -.30 despite the minimal conceptual overlap of the two topics.

Correlations between titles in raw data:

	c1	c2	c3	c4	c5	m1	m2	m3
c2	-0.19							
c3	0.00	0.00						
c4	0.00	0.00	0.47					
c5	-0.33	0.58	0.00	-0.31				
m1	-0.17	-0.30	-0.21	-0.16	-0.17			
m2	-0.26	-0.45	-0.32	-0.24	-0.26	0.67		
m3	-0.33	-0.58	-0.41	-0.31	-0.33	0.52	0.77	
m4	-0.33	-0.19	-0.41	-0.31	-0.33	-0.17	0.26	0.56

		0.02						
		-0.30	0.44					

Correlations in two dimensional space:

c2	0.91							
c3	1.00	0.91						
c4	1.00	0.88	1.00					
c5	0.85	0.99	0.85	0.81				
m1	-0.85	-0.56	-0.85	-0.88	-0.45			
m2	-0.85	-0.56	-0.85	-0.88	-0.44	1.00		
m3	-0.85	-0.56	-0.85	-0.88	-0.44	1.00	1.00	
m4	-0.81	-0.50	-0.81	-0.84	-0.37	1.00	1.00	1.00

		0.92						
		-0.72	1.00					

Figure 4. Intercorrelations among vectors representing titles (averages of vectors of the words they contain) in the original full dimensional source data of Fig. 1 and in the two-dimensional reconstruction of Fig. 3 illustrate how LSA induces passage similarity.

In the two dimensional reconstruction the topical groupings are much clearer. Most dramatically, the average r between HCI titles increases from .02 to .92. This happened, not because the HCI titles were generally similar to each other in the raw data, which they were not, but because they contrasted with the non-HCI titles in the same ways. Similarly, the correlations among the graph theory titles were re-estimated to be all 1.00, and those between the two classes of topic were now strongly negative, mean r = -.72.

Thus, SVD has performed a number of reasonable inductions; it has inferred what the true pattern of occurrences and relations must be for the words in titles if all the original data are to be accommodated in two dimensions. In this case, the inferences appear to be intuitively sensible. Note that much of the information that LSA used to infer relations among words and passages is in data about passages in which particular words did not occur. Indeed, Landauer and Dumais (1997) found that in LSA simulations of schoolchild word knowledge acquisition, about three-fourths of the gain in total comprehension vocabulary that results from reading a paragraph is indirectly inferred knowledge about words not in the paragraph at all, a result that offers an explanation of children's otherwise inexplicably rapid growth of vocabulary. A rough analogy of how this can happen is as follows. Read the following sentence:

> John is Bob's father and Mary is Ann's mother.

Now read this one:

> Mary is Bob's mother.

Because of the relations between the words mother, father, son, daughter, brother and sister that you already knew, adding the second sentence probably tended to make you think that that Bob and Ann were brother and sister, Ann the daughter of John, John the father of Ann, and Bob the son of Mary, even though none of these relations is explicitly expressed (and none follow necessarily from the presumed formal rules of English kinship naming.) The relationships inferred by LSA are also not logically defined, nor are they assumed to be consciously rationalizable as these could be. Instead, they are relations only of similarity—or of context sensitive similarity—but they nevertheless have mutual entailments of the same general nature, and also give rise to fuzzy indirect inferences that may be weak or strong and logically right or wrong.

Why, and under what circumstances should reducing the dimensionality of representation be beneficial; when, in general, will such inferences be better than the original first-order data? We hypothesize that one such case is when the original data are generated from a source of the same dimensionality and general structure as the reconstruction. Suppose, for example, that speakers or writers generate paragraphs by choosing words from a k-dimensional space in such a way that words in the same paragraph tend to be selected from nearby locations. If listeners or readers try to infer the similarity of meaning from these data, they will do better if they reconstruct the full set of relations in the same number of dimensions as the source. Among other things, given the right analysis, this will allow the system to infer that two words from nearby locations in semantic space have similar meanings even though they are never used in the same passage, or that they have quite different meanings even though they often occur in the same utterances.

The number of dimensions retained in LSA is an empirical issue. Because the underlying principle is that the original data should not be perfectly regenerated but, rather, an optimal dimensionality should be found that will cause correct induction of underlying relations, the customary factor-analytic approach of choosing a dimensionality that most parsimoniously represent the true variance of the original data is not appropriate. Instead some external criterion of validity is sought, such as the performance on a synonym test or prediction of the missing words in passages if some portion are deleted in forming the initial matrix. (See Britton & Sorrells, this issue, for another approach to determining the correct dimensions for representing knowledge.)

Finally, the measure of similarity computed in the reduced dimensional space is usually, but not always, the cosine between vectors. Empirically, this measure tends to work well, and there are some weak theoretical grounds for preferring it (see Landauer & Dumais, 1997). Sometimes we have found the additional use of the length of LSA vectors, which reflects how much was said about a topic rather than how central the discourse was to the topic, to be useful as well (see Rehder et al., this volume).

Additional detail about LSA

As mentioned, one additional part of the analysis, the data preprocessing transformation, needs to be described more fully. Before the SVD is computed, it is customary in LSA to subject the data in the raw word-by-context matrix to a two-part transformation. First, the word frequency (+ 1) in each cell is converted to its log. Second, the information-theoretic measure, entropy, of each word is computed as - p log p over all entries in its row, and each cell entry then divided by the row entropy value. The effect of this transformation is to weight each word-type occurrence directly by an estimate of its importance in the passage and inversely with the degree to which knowing that a word occurs provides information about which passage it appeared in. Transforms of this or similar kinds have long been known to provide marked improvement in information retrieval (Harman, 1986), and have been found important in several applications of LSA. The are probably most important for correctly representing a passage as a combination of the words it contains because they emphasize specific meaning-bearing words.

Readers are referred to more complete treatments for more on the underlying mathematical, computational, software and application aspects of LSA (see Berry, 1992 ; Berry, Dumais & O'Brien, 1995; Deerwester, et al., 1990; Landauer & Dumais, 1997; http://superbook.bellcore.com/~std/LSI.papers.html). On the World Wide Web site http://LSA.colorado.edu/, investigators can enter words or passages and obtain LSA based word or passage vectors, similarities between words and words, words and passages, and passages and passages, and do a few other related operations and try several prototype applications . The site offers results based on several different training corpora, such as an encyclopedia, a grade- and topic-partitioned collection of schoolchild reading, newspaper text in several languages, introductory psychology textbooks, and a small domain-specific corpus of text about the heart. To carry out LSA research based on their own training corpora, investigators will need to consult the more detailed sources (see the Appendix). Researchers should bear in mind that the LSA values given are based on samples of data and are necessarily noisy. Therefore, studies using them require the use of replicate cases and statistical treatment in a manner similar to human data.

LSA's Ability to Model Human Conceptual Knowledge

How well does LSA actually work as a representational model and measure of human verbal concepts? Its performance has been assessed more or less rigorously in several ways. We give eight examples:

(1) LSA was assessed as a predictor of query-document topic similarity judgments.

(2) LSA was assessed as a simulation of agreed upon word-word relations and of human vocabulary test synonym judgments.

(3) LSA was assessed as a simulation of human choices on subject-matter multiple choice tests.

(4) LSA was assessed as a predictor of text coherence and resulting comprehension.

(5) LSA was assessed as a simulation of word-word and passage-word relations found in lexical priming experiments.

(6) LSA was assessed as a predictor of subjective ratings of text properties, i.e. grades assigned to essays.

(7) LSA was assessed as a predictor of appropriate matches of instructional text to learners.

(8) LSA has been used with good results to mimic synonym, antonym, singular-plural and compound-component word relations, aspects of some classical word sorting studies, to simulate aspects of imputed human representation of single digits, and, in pilot studies, to replicate semantic categorical clusterings of words found in certain neuropsychological deficits (Laham, 1997b).

Kintsch (1998) has also used LSA derived meaning representations to demonstrate their possible role in construction-integration-theoretic accounts of sentence comprehension, metaphor and context effects in decision making. We will take space here to review only some of the most systematic and pertinent of these results.

LSA and information retrieval

J. R. Anderson (1990) has called attention to the analogy between information retrieval and human semantic memory processes. One way of expressing their commonality is to think of a searcher as having in mind a certain meaning, which he or she expresses in words, and the system as trying to find a text with the same meaning. Success, then, depends on the system representing query and text meaning in a manner that correctly reflects their similarity for the human. Latent Semantic Indexing (LSI; LSA's alias in this application) does this better than systems that depend on literal matches between terms in queries and documents. Its superiority can often be traced to its ability to correctly match queries to (and only to) documents of similar topical meaning when query and document use different words. In the text–processing problem to which it was first applied, automatic matching of information requests to document abstracts, SVD provides a significant improvement over prior methods. In this application, the text of the document database is first represented as a matrix of terms by documents (documents are usually represented by a surrogate such as a title, abstract and/or keyword list) and subjected to SVD, and each word and document is represented as a reduced dimensionality vector, usually with 50-400 dimensions. A query is represented as a "pseudo-document" a weighted average of the vectors of the words it contains. (A document vector in the SVD solution is also a weighted average of the vectors of words it contains, and a word vector a weighted average of vectors of the documents in which it appears.)

The first tests of LSI were against standard collections of documents for which representative queries have been obtained and knowledgeable humans have more or less exhaustively examined the whole database and judged which abstracts are and are not relevant to the topic described in each query statement. In these standard collections LSI's performance ranged from just equivalent to the best prior methods up to about 30% better. In a recent project sponsored 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, identifying proper nouns as special, differences in stop lists, and the amount of tuning 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 (essentially LSI without dimension reductions) ceteris paribus LSI was a 16% improvement (Dumais, 1994). LSI has also been used successfully to match reviewers with papers to be reviewed based on samples of the reviewers' own papers (Dumais & Nielsen, 1992), and to select papers for researchers to read based on other papers they have liked (Foltz and Dumais, 1992).

LSA and synonym tests

It is claimed that LSA, on average, represents words of similar meaning in similar ways. When one compares words with similar vectors as derived from large text corpora, the claim is largely but not entirely fulfilled at an intuitive level. Most very near neighbors (the cosine defining a near neighbor is a relative value that depends on the training database and the number of dimensions) appear closely related in some manner. In one scaling (an LSA/SVD analysis) of an encyclopedia, "physician," "patient," and "bedside" were all close to one another, cos > .5. In a sample of triples from a synonym and antonym dictionary, both synonym and antonym pairs had cosines of about .18, more than 12 times as large as between unrelated words from the same set. A sample of singular-plural pairs showed somewhat greater similarity than the synonyms and antonyms, and compound words were similar to their component words to about the same degree, more so if rated analyzable.

Nonetheless, the relationship between some close neighbors in LSA space can occasionally be mysterious (e.g., “verbally” and “sadomasochism” with a cosine of .8 from the encyclopedia space), and some pairs that should be close are not. It's impossible to say exactly why these oddities occur, but it is plausible that some words that have more than one contextual meaning receive a sort of average high-dimensional placement that out of context signifies nothing, and that many words are sampled too thinly to get well placed. It must be born in mind that most of the training corpora used to date correspond in size approximately to the printed word exposure (only) of a single average 9th grade student, and individual humans also have frequent oddities in their understanding of particular words. (Investigators who use LSA vectors should keep these factors in mind: the similarities should be expected to reflect human similarities only when averaged over many word or passages pairs of a particular type and when compared to averages across a number of people; they will not always give sensible results when applied to the particular words in a particular sentence.) It's also likely, of course, that LSA’s "bag of words" method, which ignores all syntactical, logical and nonlinguistic pragmatic entailments, sometimes misses meaning or gets it scrambled.

To objectively measure how well, compared to people, LSA captures synonymy, LSA's knowledge of synonyms was assessed with a standard vocabulary test. The 80 item test was taken from retired versions of the Educational Testing Service (ETS) Test of English as a Foreign Language (TOEFL: for which we are indebted to Larry Frase and ETS). To make these comparisons, LSA was trained by running the SVD analysis on a large corpus of representative English. In various studies, collections of newspaper text from the Associated Press news wire and Grolier's Academic American Encyclopedia (a work intended for students), and a representative collection of children’s reading[2] have

---

[2] We thank Stephen Ivens and Touchstone Applied Science Associates (TASA) of Brewster, New York for providing this valuable resource. The corpus, which was used in the production of The Educator’s Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995), consists of representative random samples of text of all kinds read by students in each grade through first year of college in the United States. In the machine-readable form in which we received it,

been used. In one experiment, an SVD was performed on text segments consisting of 500 characters or less (on average 73 words, about a paragraph's worth) taken from beginning portions of each of 30,473 articles in the encyclopedia, a total of 4.5 million words of text, roughly equivalent to what a child would have read by the end of eighth grade. This resulted in a vector for each of 60 thousand words.

The TOEFL vocabulary test consists of items in which the question part is usually a single word, and there are four alternative answers, usually single words, from which the test taker is supposed to choose the one most similar in meaning. To simulate human performance, the cosine between the question word and each alternative was calculated, and the LSA model chose the alternative closest to the stem. For six test items for which the model had never met either the stem word and/or the correct alternative, it guessed with probability .25. Scored this way, LSA got 65% correct, identical to the average score of a large sample of students applying for college entrance in the United States from non-English speaking countries.

The detailed pattern of errors of LSA was also compared to that of students. For each question a product-moment correlation coefficient was computed between (i) the cosine of the stem and each alternative and (j) the proportion of guesses for each alternative for a large sample of students (n > 1,000 for every test item). The average correlation across the 80 items was 0.70. Excluding the correct alternative, the average correlation was .44. These correlations may be thought of as between one test-taker (LSA) and group norms, which would also be much less than perfect for humans. When LSA chose wrongly and most students chose correctly, it sometimes appeared to be because LSA is more sensitive to contextual or paradigmatic associations and less to contrastive semantic or syntagmatic features. For example, LSA slightly preferred "nurse" (cos = .47) to "doctor" (cos = .41) as an associate for "physician."

---

the corpus contains approximately 11 million word tokens of text. It is one of the corpora on which LSA vectors and text similarity measures available through our Web site—http://LSA.colorado.edu—are based.

To assess the role of dimension reduction, the number of dimensions was varied from 2 to 1,032 (the largest number for which SVD was computationally feasible.) On log-linear coordinates, the TOEFL test results showed a very sharp and highly significant peak (**Figure 5**). Corrected for guessing by the standard formula ((correct - chance)/(1-chance)), LSA got 52.7% correct with 300 and 325 dimensions, 13.5% correct with just two or three dimensions. When there was no dimension reduction at all (equivalent to choosing correct answers by the correlation of transformed co-occurrence frequencies of words over encyclopedia passages), just 15.8%. At optimal dimensionality, LSA chose approximately three times as many right answers as would be obtained by ordinary first-order correlations over the input, even after a transformation that greatly improves the

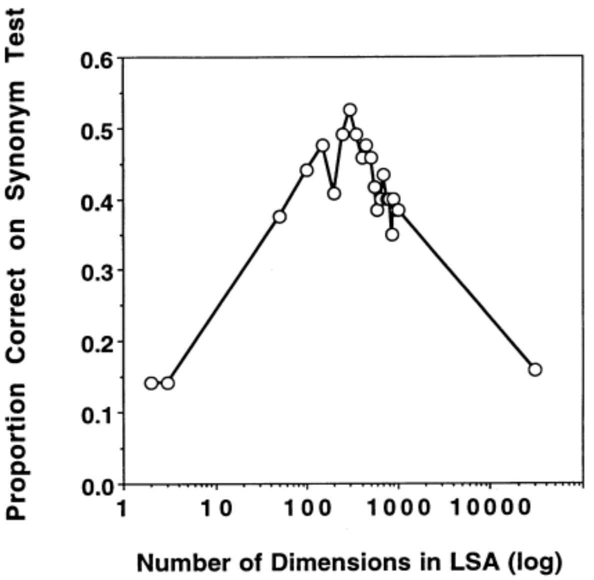

Figure 5. The effect of number of dimensions in an LSA corpus-based representation of meaning on performance on a synonym test (from ETS Test of English as a Foreign Language). The measure is the proportion of 80 multiple-choice items after standard correction for guessing. The point for the highest dimensionality is equivalent to a first-order co-occurrence correlation.

relation. This demonstrates conclusively that the LSA dimension reduction technique captures much more than mere co-occurrence (simply choosing the alternative that co-occurs with the stem in the largest number of corpus paragraphs gets only 11% right when corrected for guessing). More importantly for our argument, it implies that indirect associations or structural relations induced by analysis of the whole corpus are involved in LSA's success with individual words. Thus, correct representation of any one word depends on the simultaneous correct representation of many, perhaps all other words.

As mentioned earlier, Landauer and Dumais (1997) also estimated, by a different method, the relative direct and indirect effects of adding a new paragraph to LSA's "experience". For example, at a point in LSA's learning roughly corresponding to the amount of text read by late primary school, an imaginary test of all words in the language—the model's imputed total recognition vocabulary—gains about three times as much knowledge about words <u>not in</u> the new paragraph as about words actually contained in the paragraph.

Landauer and Dumais (1997) also found that the rate of gain in vocabulary by LSA was approximately equal to the rate of gain of "known", as compared to morphologically inferred, words empirically estimated by Anglin (1995) and others for primary school children.

<u>**Simulating word sorting and relatedness judgments**</u>

Recently, Laham and Landauer explored the relation between LSA and human lexical semantic representations further by simulating a classic word sorting study by Anglin (1970). In Anglin's experiments third and fourth grade children and adults were given sets of selected words to sort by meaning into as many piles as they wished. The word sets contained subsets of nouns, verbs, prepositions and adjectives, and within each subset there were words taken from common conceptual hierarchies, such as <u>boy, girl, horse, flower,</u> among which clustering could reveal the participant's tendency to use abstract versus concrete similarity relations. Anglin measured the semantic similarity of every pair of words by the proportion of subjects who put them in the same pile. He found that parts of speech clustered moderately in both child and adult sets, and, confirming the hypothesis behind the study, that adults showed more evidence of use of abstract categories than did children.

Laham and Landauer measured the similarity between the same word pairs by cosines based on 5 grade-partitioned scalings of samples of schoolchild reading—3rd, 6th, 9th, 12th grade and college.[3] For each scaling, the cosine between each word pair in the set (20 words for 190 comparisons) was calculated. The overall correlation of the LSA estimates and the grouped human data, for both child and adult, rose as the number of documents included in the scaling rose. Using the third grade scaling, the correlation between the LSA estimates and the child data was .50, with the adult data .35. Using the college level scaling the correlation between LSA estimates and the child data was .61, with adults .50. The correlation coefficients between LSA estimates and human data showed a monotonic linear rise as the grade level (and number of documents known to LSA) increased.

LSA exhibited differences in similarities across degrees of abstraction much like those found by Anglin; for the third grade scaling, the average correlations in patterns across means for the comparable levels within each part-of-speech class $\underline{r} = .80$ with children and $\underline{r} = .75$ with adults, for the college level scaling $\underline{r} = .90$ with children and $\underline{r} = .90$ with adults . The correlation between the adult and child patterns was .95. The LSA measure did not separate word classes nearly as strongly as did the human data, nor did it as clearly distinguish within part-of-speech from between part-of-speech comparisons. For the third grade scaling, the overall (N = 190) average cosine =.13, the average within part-of-speech cosine (N = 41) = .15 and the average between part-of-speech cosine (N = 149) = .13. The college level scaling showed stronger similarities within class with the overall

---

[3] See previous footnote.

average cosine =.19, the average within part-of-speech cosine = .23 and the average between part-of-speech cosine = .17.

As in the vocabulary acquisition simulations, it appears that the relations obtained from a corpus of small size relative to a typical adult's cumulative language exposure resemble children somewhat more than adults. LSA's weak reflection of word class in this rather small sample of data would appear to confirm the expectation that the lack of word order information in its input data along with the use of fairly large passages as the context units prevents it from inducing grammatical relations among words. (Wolfe et al., 1998/this issue, reports further word sorting results. Also compare Burgess et al, 1998/this issue.)

Simulating subject-matter knowledge

In three investigations by Foltz and by Laham and Landauer (Landauer, Foltz, & Laham, 1998) to be reported fully elsewhere, LSA has been trained on the text of introductory psychology textbooks, then tested with multiple choice tests provided by the textbook publishers. LSA performed well above chance in all cases, and in all cases did significantly better on questions rated "easy" than on ones rated "difficult", and on items classified as "factual" than on ones classified as "conceptual" by their authors. On questions used in university introductory psychology course exams given at New Mexico State University and the University of Colorado, Boulder, LSA scored significantly worse than class averages, but in every case did well enough to receive a passing grade according to the class grading scheme.

In related work, Foltz, Britt and Perfetti (1996) used LSA to model the knowledge structures of both expert and novice subjects who had read a large number of documents on the history of the Panama canal. After reading the documents, subjects made judgments of the relatedness of 120 pairs of concepts that were mentioned in the documents. Based on an LSA scaling of the documents, the cosines between the concepts were used to estimate the relatedness of the concept pairs. The LSA predictions correlated significantly with the subjects, with the correlation stronger to that of the experts in the domain ($\underline{r} = 0.41$) than that of the novices ($\underline{r} = 0.36$). (Note again that two human ratings would also not correlated perfectly.) An analysis of where LSA's predictions deviated greatly from that of the humans indicated that LSA tended to underpredict more global or situational relationships that were not directly discussed in the text but would be common historical knowledge of any undergraduate. Thus in this case the limitation on LSA's predictions may simply be due to training only on a small set of documents rather than on a larger set that would capture a richer representation of history.

Simulating semantic priming

Landauer and Dumais (1997) report an analysis in which LSA was used to simulate a lexical semantic priming study by Till, Mross and Kintsch (1988), in which people were presented visually with one or two sentence passages that ended in an obviously polysemous word. After varying onset delays, participants made lexical decisions about words related to the homographic word or to the overall meaning of the sentence. In paired passages, each homographic word’s meaning was biased in two different ways judged to be related to two corresponding different target words. There were two additional target words not in the passages or obviously related to the polysemous word but judged to be related to the overall meaning or “situation model” that people would derive from the passage. Here is an example of two passages and their associated target words, along with a representative control word used to establish a baseline.

“The townspeople were amazed to find that all the buildings had collapsed except the mint.”
“Thinking of the amount of garlic in his dinner, the guest asked for a mint.”

Target words: money, candy, earthquake, breath
Unrelated control word: ground

In the Till et al. study, target words related to both senses of the homographic words were correctly responded to faster than unrelated control words if presented within 100 ms after the homograph. If delayed by 300 ms, only the context-appropriate associate was primed. At a one second delay, the so-called inference words were also primed. In the LSA simulation, the cosines between the polysemic word and its two associates were computed to mimic the expected initial priming. The cosine between the two associates of the polysemic word and the sentence up to the last word preceding it were used to mimic contextual disambiguation of the homographs. The cosine between the entire passages and the inference words were computed to emulate the contextual comprehension effect on their priming.

Table 1 shows the average results over all 27 passage pairs, with one of the above example passages shown again to illustrate the conditions simulated. The values given are the cosines between the word or passage and the target words. The pattern of LSA similarity relations corresponds almost perfectly with the pattern of priming results; the differences corresponding to differences observed in the priming data are all significant at p < .001, and have effect sizes comparable to those in the priming study.

The import of this result is that LSA again emulated a human behavioral relation between words and multi-word passages, and did so while representing passages simply as the vector average of their contained words. (Steinhart, 1995, obtained similar results with different words and passages.) It is surprising and important that such simple representations of whole utterances, ones that ignore word order, sentence structure, and non-linear word-word interactions, can correctly predict human behavior based on passage meaning. However, this is the second example of this property—query-abstract and abstract-abstract similarity results being the first—and there have subsequently been several more. These findings begin to suggest that word choice alone has a much more dominant role in the expression of meaning than has previously been credited (see Landauer, Laham and Foltz, 1997).

Table 1

LSA Simulation of Till, Mross, & Kintsch (1988) Priming Study.

Mint:

Money	Candy	Ground
.21	.20	.07
Thinking amount garlic dinner guest asked:		
Money	Candy	
.15	.21	
		Ground
Earthquake	Breath	.15
.14	.21	

Note. LSA = Latent Semantic Analysis.

Of course, LSA as currently constituted contains no model of the temporal dynamics of discourse comprehension. To fit the temporal findings of the Till et al. experiment one would need to assume that the combining (averaging) of word vectors into a single vector to represent the whole passage takes about a second, and that partial progress of the combining mechanism accounts for the order and times at which the priming changes occur. We hope eventually to develop dynamic LSA-based models of the word combining mechanism by which sentence and passage comprehension is accomplished. Such models will presumably incorporate LSA word representations into processes like those posited in Construction-Integration (Kintsch, 1988) or other spreading activation theories. An example of such a model would be to first compute the vector of each word, then the average vector for the two most similar words, and so forth. It seems likely that such a model would prove too simple. However, the research strategy behind the LSA effort would dictate trying the simplest models first and then complicating them, for example in the direction of the full-blown CI construction and iterative constraint satisfaction mechanisms, or even to models including hierarchical syntactic structure (presumably, automatically induced), only if and as found necessary.

Assigning holistic quality scores to essay test answers

In another set of studies to be published elsewhere by Landauer, Laham and Foltz (1998), LSA has been used to assign holistic quality scores to written answers to essay questions. Five different methods have been tried, all with good success. In all cases an LSA space was first constructed based either on the instructional text read by students or on similar text from other sources, plus the text of student essays. In Method 1, a sample of essays was first graded by instructors, then the cosine (or other LSA-based similarity and quantity measures, or both) between each ungraded essay and each pre-graded essay was computed, and the new essay assigned the average of a small set of closely similar ones, weighted by their similarity.

In Method 2, a pre-existing exemplary text on the assigned topic, one written by an instructor or expert author, was used as a standard, and the student essay score was computed as its LSA cosine with the standard. In the Method 3, the cosine between each sentence of a standard text from which the students had presumably learned the material being tested and each sentence of a student’s answer was first computed. The maximum cosine for each source text component was found among the sentences of the student essay, and these cumulated to form a total score. In a variant of the third method, Method 4 computed and cumulated the cosines between each sentence in a student's essay and a set of sentences from the original text that the instructor thought were important.

In Method 5, only the essays themselves were used. The matrix of distances (1-cosine) between all essays was "unfolded" to the single dimension that best reconstructed all the distances, and the point of an essay along this dimension taken as the measure of its quality. This assumes that the most important dimension of difference among a set of essay exams on a given topic is their global quality.

All five methods provided the basis of scores that correlated approximately as well with expert assigned scores as such scores correlated with each other, sometimes slightly less well, on average somewhat better. In one set of studies (Laham, 1997a), method one was applied to a total of eight exams ranging in topic from heart anatomy and physiology, through psychological concepts, to American history, current social issues and marketing problems. A meta-analysis found that LSA correlated significantly better with individual expert graders (from ETS or other professional organization or course instructors) than one expert correlated with another.

Because these results show that human judgments about essay qualities are no more reliable than LSA's, they again suggest that the holistic semantic representation of a passage relies primarily on word choice and surprisingly little on properties whose transmission necessarily requires the use of syntax. This is good news for the practical application of LSA to many kinds of discourse processing research, but is counter-intuitive and at odds with the usual assumptions of linguistic and psycholinguistic theories of meaning and comprehension, so it should be viewed with caution until further research is done (and, of course, with reservations until the details of the studies have been published.)

LSA and Text Comprehension

This application of LSA is described in papers in this volume, so we will mention the results only briefly to round out our survey of evidence regarding the quality of LSA's simulation of human meaning-based performance. Kintsch and his colleagues (e.g. van Dijk & Kintsch, 1983; Kintsch & Vipond, 1979; McNamara, Kintsch, Songer & Kintsch, 1996) have developed methods for representing text in a propositional language and have used it to analyze the coherence of discourse. They have shown that the comprehension of text depends heavily on its coherence, as measured by the overlap between the arguments in propositions. In a typical propositional calculation of coherence, a text must first be propositionalized by hand. This has limited research to small samples of text and has inhibited its practical application to composition and instruction. Foltz, Kintsch, and Landauer (1993, this issue; Foltz, 1996) have applied LSA to the task. LSA can make automatic coherence judgments by computing the cosine from one sentence or passage and the following one. In one case, analysis of the coherence between a set of sentences about the heart, the LSA measure predicted comprehension scores extremely well, r= .93. As will be discussed in the article in this volume, the general approach of using LSA for computing textual coherence also permits an automatic characterization of places in a text where the coherence breaks down, as well as a measure of how semantic content changes across a text.

Predicting learning from text

As reported in some detail in two of the succeeding articles in this issue, Kintsch, Landauer and colleagues (Rehder et al.; Wolfe et al.; this issue) have begun to use LSA to match students with text at the optimal level of conceptual complexity for learning. Earlier work by Kintsch and his collaborators (see Kintsch, 1994; McNamara, Kintsch, Butler-Songer and Kintsch, 1996) has shown 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, requiring too little new knowledge construction (a phenomenon we call “the Goldilocks principle”). LSA has been used to characterize both the knowledge of an individual student before and after reading a particular text and the knowledge conveyed by that text. These studies and their results are described in detail in articles hereafter. It is shown that choosing between instructional texts of differing sophistication by the LSA relation between a short student essay and the text can significantly increase the amount learned. In addition, analytic methods are developed by which not only the similarity between two or more texts, but their relative positions along some important underlying conceptual continuum, such as level of sophistication or relevance to a particular topic, can be measured.

Summary and some caveats

It is clear enough from the conjunction of all these formal and informal results that LSA is able to capture and represent significant components of the lexical and passage meanings evinced in judgment and behavior by humans. The following papers exploit this ability in interesting and potentially useful ways that simultaneously provide additional demonstrations and tests of the method and its underlying theory. However, as mentioned briefly above, it is obvious that LSA lacks important cognitive abilities that humans use to construct and apply knowledge from experience, in particular the ability to use detailed and complex order information such as that expressed by syntax and used in logic. It also lacks, of course, a great deal of the important raw experience, both linguistic and otherwise, on which human knowledge is based. While we are impressed by LSA's current power to mimic aspects of lexical semantics and psycholinguistic phenomena, we believe that its validity as a model or measure of human cognitive processes or their products should not be oversold. When applied in detail to individual cases of word pair relations or sentential meaning construal it often goes awry when compared to our intuitions. In general, it performs best when used to simulate average results over many cases, suggesting either that, so far at least, it is capturing statistical regularities that emerge from detailed processes rather than the detailed processes themselves, or that the corpora and, perhaps, the analysis methods, used to date have been imperfect.

On the other hand, the success of LSA as a theory of human knowledge acquisition and representation should also not be underestimated. It is hard to imagine that LSA could have simulated the impressive range of meaning-based human cognitive phenomena that it has unless it is doing something analogous to what humans do. No previous theory in linguistics, psychology or artificial intelligence research has ever been able to provide a rigorous computational simulation that takes in the very same data from which humans learn about words and passages and produces a representation that gives veridical simulations of a wide range of human judgments and behavior. While it seems highly doubtful that the human brain uses the same mathematical algorithms as LSA/SVD, it seems almost certain that the brain uses as much analytic power as LSA to transform its temporally local experiences into global knowledge. The present theory clearly does not account for all aspects of knowledge and cognition, but it offers a potential path for development of new accounts of mind that can be stated in mathematical terms rather than imprecise mentalistic primitives and whose empirical implications can be derived analytically or by computations on bodies of representative data rather than by verbal argument.

In future research we hope to see both improvements in LSA's experience base from analysis of larger and more representative corpora of both text and spoken language—and perhaps, if a way can be found, by adding representations of experience of other kinds—and the provision of a compatible process model of online discourse comprehension by which both its input of experience and its application of constructed knowledge will better reflect the complex ways in which humans combine word meanings dynamically. As suggested above, one promising approach to the latter goal is to combine LSA word and episode representation with the Construction-Integration theory's mechanisms for discourse comprehension, a strategy that Walter Kintsch illustrates in a forthcoming book (Kintsch, in press.) Other avenues of potential improvement involve the representation of word order in the input data for LSA, following the example of the work reported in Burgess and Lund (this volume).

Meanwhile, it needs keeping in mind that the applications of LSA recounted in the following articles are all based on its current formulation and based on varying training corpora that are all smaller and less representative of relevant human experience than one would wish. Part of the problem of non-optimal corpora is due simply to the current unavailability and difficulty of constructing large general or topically relevant text samples that approximate what a variety of individual learners would have met. But another is due to current computational limitations. LSA became practical only when computational power and algorithm efficiency improved sufficiently to support SVD of thousands of words-by-thousands of contexts matrices; it is still impossible to perform SVD on the hundreds of thousands by tens of millions matrices that would be needed to truly represent the sum of an adult's language exposure. It also needs noting that is still early days for LSA and that many details of its implementation, such as the preprocessing data transformation used and the method for choosing dimensionality, even the underlying statistical model, will undoubtedly undergo changes.

Thus in reading the following articles, or in considering the application of LSA to other problems, one should not think of LSA as a fixed mechanism or its representations as fixed quantities, but rather, as evolving approximations.

REFERENCES

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anglin, J. M. (1970) The growth of word meaning. Cambridge, MA.: MIT Press.

Anglin, J. M., Alexander, T. M., & Johnson, C. J. (1996). Word learning and the growth of potentially knowable vocabulary. Submitted for publication.

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

Berry, M. W., Dumais, S. T. and O'Brien, G.W. (1995) Using linear algebra for intelligent information retrieval. SIAM: Review, 37, 573-595.

Britton, B. K. & Sorrells, R. C. (1998/this issue). Thinking about knowledge learned from instruction and experience: Two tests of a connectionist model. Discourse Processes, 25, 131-177.

Burgess, C., Livesay, K. & Lund, K. (1998/this issue). Explorations in context space: Words, sentences, discourse. Discourse Processes, 25, 211-257.

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.), The Second Text Retrieval Conference (TREC2) (National Institute of Standards and Technology Special Publication 500-215, pp. 105-116).

Dumais, S. T. & Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In N. Belkin, P. Ingwesen, & A. M. Pejtersen (Eds.) Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, Association for Computing Machinery.

Foltz, P. W. (1996) Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers. 28, 197-202.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996) Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. Cottrell (Ed.) Proceedings of the 18th Annual Cognitive Science Conference. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P. W. & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. Communications of the ACM, 35, 51-60.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1993, July). An analysis of textual coherence using Latent Semantic Indexing. Paper presented at the meeting of the Society for Text and Discourse, Boulder, CO.

Harman, D. (1986). An experimental study of the factors important in document ranking. In Association for Computing Machinery Conference on Research and Development in Information Retrieval. Association for Computing Machinery.

Kintsch, W. (1988) The role of knowledge in discourse comprehension construction-integration model. Psychological Review, 163-182.

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

Kintsch, W. (1998) Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Eds.), Perspectives on Memory Research. Hillsdale, NJ: Erlbaum.

Laham, D. (1997a). Automated holistic scoring of the quality of content in directed student essays using Latent Semantic Analysis. Unpublished master's thesis, University of Colorado, Boulder.

Laham, D. (1997b). Latent Semantic Analysis Approaches to Categorization. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th Annual Conference of the Cognitive Science Society (p. 979). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Landauer, T.K. and Dumais, S.T. (1994). Latent semantic analysis and the measurement of knowledge. In R. M. Kaplan and J. C. Burstein (Eds) Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education. Princeton, Educational Testing Service.

Landauer, T. K., & Dumais, S. T. (1996). How come you know so much? From practical problems to new memory theory. In D. J. Hermann, C. McEvoy, C. Hertzog, P. Hertel, & M. K. Johnson (Eds.), Basic and applied memory research: Vol. 1. Theory in context (pp. 105-126). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-140

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Latent Semantic Analysis passes the test: knowledge representation and multiple-choice testing. Unpublished manuscript.

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Computer-based grading of the conceptual content of essays. Unpublished manuscript.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E., (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412-417). Mawhwah, NJ: Erlbaum.

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

McNamara, D. S., Kintsch, E., Songer B. N., & Kintsch, W. (1996) Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1, 1-43.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998/this issue). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. Discourse Processes, 25, 337-354.

Steinhart, D. J. (1996). Resolving Lexical ambiguity: Does context play a role? Unpublished master's thesis, University of Colorado, Boulder.

Till, R. E. , Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in discourse context. Memory and Cognition, 16, 283-298.

van Dijk, T. A., & Kintsch, W. (1983). Strategies of Discourse Comprehension. New York: Academic Press.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998/this issue). Learning from text: Matching readers and text by Latent Semantic Analysis. Discourse Processes, 25, 309-336.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educator's word frequency guide. Brewster, NY: Touchstone Applied Science Associates.

Appendix

The latest information and applications of LSA can be found at our website:

http://LSA.colorado.edu/

This website is organized into three content areas, Information, Demonstrations, and Applications. The Information section contains additional papers, links, and other pertinent information on LSA.

The Demonstrations section currently includes examples of essay scoring and matching learners to text. The matching application allows you to explore the use of LSA as a tool for selecting texts that will augment learning. The demonstration shows how LSA might be used to select a text about the heart based on the knowledge demonstrated in a short essay. The returned text should be understandable to the reader as well as help him or her learn something new.

The Applications section permits you to select an available LSA semantic space and run some comparison experiments on text you provide. Each application consists of a form where you are to include the text(s) that you want to make LSA comparisons with (as well as a number of options). After you submit the form, the LSA programs will make the desired comparisons and return the results to a new web page. You can save the results using your browser's Save Frame menu.

Author Note

Darrell Laham and Thomas K Landauer, Department of Psychology, University of Colorado, Boulder. Peter W. Foltz, Department of Psychology, New Mexico State University.

This research was supported in part by a contract from ARPA-CAETI to T. Landauer and W. Kintsch.

Correspondence concerning this article should be addressed to Thomas K Landauer, Department of Psychology, Campus Box 345, University of Colorado, Boulder, CO, 80309. Electronic mail may be sent via Internet to landauer@psych.colorado.edu.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes,* **25**, 259-284.