Psychological Review
1997, Vol. 104, No. 2, 211-240

Copyright 1997 by the American Psychological Association, Inc.
0033-295X/97/$3.00

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

Susan T. Dumais
Bellcore

How do people know as much as they do with as little information as they get? The problem takes many forms; learning vocabulary from text is an especially dramatic and convenient case for research. A new general theory of acquired similarity and knowledge representation, latent semantic analysis (LSA), is presented and used to successfully simulate such learning and several other psycholinguistic phenomena. By inducing global knowledge indirectly from local co-occurrence data in a large body of representative text, LSA acquired knowledge about the full vocabulary of English at a comparable rate to schoolchildren. LSA uses no prior linguistic or perceptual similarity knowledge; it is based solely on a general mathematical learning method that achieves powerful inductive effects by extracting the right number of dimensions (e.g., 300) to represent objects and contexts. Relations to other theories, phenomena, and problems are sketched.

Prologue

> "How much do we know at any time? Much more, or so I believe, than we know we know!"
>
> —Agatha Christie, *The Moving Finger*

A typical American seventh grader knows the meaning of 10–15 words today that she did not know yesterday. She must have acquired most of them as a result of reading because (a) the majority of English words are used only in print, (b) she already knew well almost all the words she would have encountered in speech, and (c) she learned less than one word by direct instruction. Studies of children reading grade-school text find that about one word in every 20 paragraphs goes from wrong to right on a vocabulary test. The typical seventh grader would have read less than 50 paragraphs since yesterday, from which she should have learned less than three new words. Apparently, she mastered the meanings of many words that she did not encounter. Evidence for all these assertions is given in detail later.

This phenomenon offers an ideal case in which to study a problem that has plagued philosophy and science since Plato 24 centuries ago, the fact that people have much more knowledge than appears to be present in the information to which they have been exposed. Plato's solution, of course, was that people must come equipped with most of their knowledge and need only hints and contemplation to complete it.

In this article we suggest a very different hypothesis to explain the mystery of excessive learning. It rests on the simple notion that some domains of knowledge contain vast numbers of weak interrelations that, if properly exploited, can greatly amplify learning by a process of inference. We have discovered that a very simple mechanism of induction, the choice of the correct dimensionality in which to represent similarity between objects and events, can sometimes, in particular in learning about the similarity of the meanings of words, produce sufficient enhancement of knowledge to bridge the gap between the information available in local contiguity and what people know after large amounts of experience.

Overview

In this article we report the results of using latent semantic analysis (LSA), a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism, to analyze a large corpus of natural text and generate a representation that captures the similarity of words and text passages. The model's resulting knowledge was tested with a standard multiple-choice synonym test, and its learning power compared to the rate at which school-aged children improve their performance on similar tests as a result of reading. The model's improvement per paragraph of encountered text approximated the natural rate for schoolchildren, and most of its acquired knowledge was attributable to indirect inference rather than direct co-occurrence relations. This result can be interpreted in at least two ways. The more conservative interpretation is that it shows that with the right analysis a substantial portion of the information needed to answer common vocabulary test questions can be inferred from the contextual statistics of usage alone. This is not a trivial conclusion. As we alluded to earlier

Thomas K Landauer, Institute of Cognitive Science, University of Colorado at Boulder; Susan T. Dumais, Information Science Research Department, Bellcore, Morristown, New Jersey.

We thank Karen Lochbaum for valuable help in analysis; George Furnas for early ideas and inspiration; Peter Foltz, Walter Kintsch, and Ernie Mross for unpublished data; and for helpful comments on the ideas and drafts, we thank, in alphabetic order, Richard Anderson, Doug Carroll, Peter Foltz, George Furnas, Walter Kintsch, Lise Menn, and Lynn Streeter.

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

and elaborate later, much theory in philosophy, linguistics, artificial intelligence research, and psychology has supposed that acquiring human knowledge, especially knowledge of language, requires more specialized primitive structures and processes, ones that presume the prior existence of special foundational knowledge rather than just a general purpose analytic device. This result questions the scope and necessity of such assumptions. Moreover, no previous model has been applied to simulate the acquisition of any large body of knowledge from the same kind of experience used by a human learner.

The other, more radical, interpretation of this result takes the mechanism of the model seriously as a possible theory about all human knowledge acquisition, as a homologue of an important underlying mechanism of human cognition in general. In particular, the model employs a means of induction—dimension optimization—that greatly amplifies its learning ability, allowing it to correctly infer indirect similarity relations only implicit in the temporal correlations of experience. The model exhibits humanlike generalization that is based on learning and that does not rely on primitive perceptual or conceptual relations or representations. Similar induction processes are inherent in the mechanisms of certain other theories (e.g., some associative, semantic, and neural network models). However, as we show later, substantial effects arise only if the body of knowledge to be learned contains appropriate structure and only when a sufficient—possibly quite large—quantity of it has been learned. As a result, the posited induction mechanism has not previously been credited with the significance it deserves or exploited to explain the many poorly understood psychological phenomena to which it may be germane. The mechanism lends itself, among other things, to a deep reformulation of associational learning theory that appears to offer explanations and modeling directions for a wide variety of cognitive phenomena. One set of phenomena that we discuss later in detail, along with some auxiliary data and simulation results, is contextual disambiguation of words and passages in text comprehension.

Because readers with different theoretical interests may find these two interpretations differentially attractive, we have followed a slightly unorthodox manner of exposition. Although we later present a general theory, or at least the outline of one, that incorporates and fleshes out the implications of the inductive mechanism of the formal model, we have tried to keep this development somewhat independent of the report of our simulation studies. That is, we eschew the conventional stance that the theory is primary and the simulation studies are tests of it. Indeed, the historical fact is that the mathematical text analysis technique came first, as a practical expedient for automatic information retrieval, the vocabulary acquisition simulations came next, and the theory arose last, as a result of observed empirical successes and discovery of the unsuspectedly important effects of the model's implicit inferential operations.

The Problem of Induction

One of the deepest, most persistent mysteries of cognition is how people acquire as much knowledge as they do on the basis of as little information as they get. Sometimes called "Plato's problem" or "the poverty of the stimulus," the question is how observing a relatively small set of events results in beliefs that are usually correct or behaviors that are usually adaptive in a large, potentially infinite variety of situations. Following Plato, philosophers (e.g., Goodman, 1972; Quine, 1960), psychologists (e.g., Shepard, 1987; Vygotsky, 1968), linguists (e.g., Chomsky, 1991; Jackendoff, 1992; Pinker, 1990), computation scientists (e.g., Angluin & Smith, 1983; Michaelski, 1983) and combinations thereof (Holland, Holyoak, Nisbett, & Thagard, 1986) have wrestled with the problem in many guises. Quine (1960), following a tortured history of philosophical analysis of scientific truth, has called the problem "the scandal of induction," essentially concluding that purely experience-based objective truth cannot exist. Shepard (1987) has placed the problem at the heart of psychology, maintaining that a general theory of generalization and similarity is as necessary to psychology as Newton's laws are to physics. Perhaps the most well-advertised examples of the mystery lie in the acquisition of language. Chomsky (e.g., Chomsky, 1991) and followers assert that a child's exposure to adult language provides inadequate evidence from which to learn either grammar or lexicon. Gold, Osherson, Feldman, and others (see Osherson, Weinstein, & Stob, 1986) have formalized this argument, showing mathematically that certain kinds of languages cannot be learned to certain criteria on the basis of finite data. The puzzle presents itself with quantitative clarity in the learning of vocabulary during the school years, the particular case that we address most fully in this article. Schoolchildren learn to understand words at a rate that appears grossly inconsistent with the information about each word provided by the individual language samples to which they are exposed and much faster than they can be made to by explicit tuition.

Recently Pinker (1994) has summarized the broad spectrum of evidence on the origins of language—in evolution, history, anatomy, physiology, and development. In accord with Chomsky's dictum, he concludes that language learning must be based on a very strong and specific innate foundation, a set of general rules and predilections that need parameter setting and filling in, but not acquisition as such, from experience. Although this "language instinct" position is debatable as stated, it rests on an idea that is surely correct, that some powerful mechanism exists in the minds of children that can use the finite information they receive to turn them into competent users of human language. What we want to know, of course, is what this mechanism is, what it does, how it works. Unfortunately the rest of the instinctivist answers are as yet of limited help. The fact that the mechanism is given by biology or that it exists as an autonomous mental or physical "module" (if it does), tells us next to nothing about how the mind solves the basic inductive problem.

Shepard's (1987) answer to the induction problem in stimulus generalization is equally dependent on biological givens, but offers a more precise description of some parts of the proposed mechanism. He has posited that the nervous system has evolved general functional relations between monotone transductions of perceptual values and the similarity of central interpretive processes. On average, he has maintained, the similarities generated by these functions are adaptive because they predict in what situations—*consequential regions* in his terminology—the same behavioral cause–effect relations are likely to hold. Shepard's mathematical laws for stimulus generalization are empiri-

cally correct or nearly so for a considerable range of low-dimensional perceptual continua and for certain functions computed on behaviorally measured relations such as choices between stimuli or judgments of similarity or inequality on some experiential dimension. However, his laws fall short of being able to predict whether cheetahs are considered more similar to zebras or tigers, whether friendship is thought to be more similar to love or hate, and are mute, or at least very incomplete, on the similarity of the meanings of the words *cheetah, zebra, tiger, love, hate,* and *pode.* Indeed, it is the generation of psychological similarity relations based solely on experience and the achievement of bridging inferences from experience about cheetahs and friendship to behavior about tigers and love and from hearing conversations about one to knowledge about the other that pose the most difficult and tantalizing puzzle.

Often the cognitive aspect of the induction puzzle is cast as the problem of categorization, of finding a mechanism by which a set of stimuli, words, or concepts (*cheetahs, tigers*) come to be treated as the same for some purposes (running away from, or using metaphorically to describe a friend or enemy). The most common attacks on this problem invoke similarity as the underlying relation among stimuli, concepts, or features (e.g., Rosch, 1978; Smith & Medin, 1981; Vygotsky, 1968). But as Goodman (1972) has trenchantly remarked, "similarity is an impostor," at least for the solution of the fundamental problem of induction. For example, the categorical status of a concept is often assumed to be determined by similarity to a prototype, or to some set of exemplars (e.g., Rosch, 1978; Smith & Medin, 1981). Similarity is either taken as primitive (e.g., Posner & Keele, 1968; Rosch, 1978) or as dependent on shared component features (e.g., Smith & Medin, 1981; Tversky, 1977; Tversky & Gati, 1978). But this throws us into an unpleasant regress: When is a feature a feature? Do bats have wings? When is a wing a wing? Apparently, the concept *wing* is also a category dependent on the similarity of features. Presumably, the regress ends when it grounds out in the primitive perceptual relations assumed, for example, by Shepard's theory. But only some basic perceptual similarities are relevant to any feature or category, others are not; a wing can be almost any color. The combining of disparate things into a common feature identity or into a common category must very often depend on experience. How does that work? Crisp categories, logically defined on rules about feature combinations, such as those often used in category learning, probability estimation, choice and judgment experiments, lend themselves to acquisition by logical rule–induction processes, although whether such processes are what humans always or usually use is questionable (Holland, Holyoak, Nisbett, & Thagard, 1986; Medin, Goldstone, & Gentner, 1993; Murphy & Medin, 1985; Smith & Medin, 1981). Surely, the natural acquisition of fuzzy or probabilistic features or categories relies on some other underlying process, some mechanism by which experience with examples can lead to treating new instances more or less equivalently, some mechanism by which common significance, common fate, or common context of encounter can generate acquired similarity. We seek a mechanism by which the experienced and functional similarity of concepts—especially complex, largely arbitrary ones, such as the meaning of *concept, component,* or *feature,* or, perhaps, the component features of which concepts might consist—are created from an interaction of experience with the logical (or mathematical or neural) machinery of mind.

In attempting to explain the astonishing rate of vocabulary learning—some 7–10 words per day—in children during the early years of preliterate language growth, theorists such as Carey (1985), Clark (1987), Keil (1989), and Markman (1994) have hypothesized constraints on the assignment of meanings to words. For example it has been proposed that early learners assume that most words are names for perceptually coherent objects, that any two words usually have two distinct meanings, that words containing common sounds have related meanings, that an unknown speech sound probably refers to something for which the child does not yet have a word, and that children obey certain strictures on the structure of relations among concept classes. Some theorists have supposed that the proposed constraints are biological givens, some have supposed that they derive from progressive logical derivation during development, some have allowed that constraints may have prior bases in experience. Many have hedged on the issue of origins, which is probably not a bad thing, given our state of knowledge. For the most part, proposed constraints on lexicon learning have also been described in qualitative mentalistic terminology that fails to provide entirely satisfying causal explanations: Exactly how, for example does a child apply the idea that a new word has a new meaning?

What all modern theories of knowledge acquisition (as well as Plato's) have in common is the postulation of constraints that greatly (in fact, infinitely) narrow the solution space of the problem that is to be solved by induction, that is, by learning. This is the obvious, indeed the only, escape from the inductive paradox. The fundamental notion is to replace an intractably large or infinite set of possible solutions with a problem that is soluble on the data available. So, for example, if biology specifies a function on wavelength of light that is assumed to map the difference between two objects that differ only in color onto the probability that doing the same thing with them will have the same consequences, then a bear need sample only one color of a certain type of berry before knowing which others to pick.

There are several problematical aspects to constraint-based resolutions of the induction paradox. One is whether a particular constraint exists as supposed. For example, is it true that young children assume that the same object is given only one name, and if so is the assumption correct about the language to which they are exposed? (It is not in adult English usage; ask 100 people what to title a recipe or name a computer command, and you will get almost 30 different answers on average—see Furnas, Landauer, Gomez, & Dumais, 1983, 1987). These are empirical questions, and ones to which most of the research in early lexical acquisition has been addressed. One can also wonder about the origin of a particular constraint and whether it is plausible to regard it as a primitive process with an evolutionary basis. For example, most of the constraints proposed for language learning are very specific and relevant only to human language, making their postulation consistent with a very strong instinctive and modular view of mental processes.

The existence and origin of particular constraints is only one part of the problem. The existence of some set of constraints is a logical necessity, so that showing that some exist is good but not nearly enough. We also need to know whether a particular

set of constraints is logically and pragmatically sufficient, that is, whether the problem space remaining after applying them is soluble. For example, suppose that young children do, in fact, assume that there are no synonyms. How much could that help them in learning the lexicon from the language to which they are exposed? Enough? Indeed, that particular constraint leaves the mapping problem potentially infinite; it could even exacerbate the problem by tempting the child to assign too much or the wrong difference to *our dog, the collie,* and *Fido.* Add in the rest of the constraints that have been proposed: Enough now?

How can one determine whether a specified combination of constraints would solve the problem, or perhaps better, determine how much of the problem it would solve? We believe that the best available strategy is to specify a concrete computational model embodying the proposed constraints and to simulate as realistically as possible its application to the acquisition of some measurable and interesting properties of human knowledge. In particular, with respect to constraints supposed to allow the learning of language and other large bodies of complexly structured knowledge, domains in which there are very many facts each weakly related to very many others, effective simulation may require data sets of the same size and content as those encountered by human learners. Formally, that is because weak local constraints can combine to produce strong inductive effects in aggregate. A simple analog is the familiar example of a diagonal brace to produce rigidity in a structure made of three beams. Each connection between three beams can be a single bolt. Two such connections exert no constraint at all on the angle between the beams. However, when all three beams are so connected, all three angles are completely specified. In structures consisting of thousands of elements weakly connected (i.e., constrained) in hundreds of different ways (i.e., in hundreds of dimensions instead of two), the effects of constraints may emerge only in very large, naturally generated ensembles. In other words, experiments with miniature or concocted subsets of language experience may not be sufficient to reveal or assess the forces that hold conceptual knowledge together. The relevant quantitative effects of such phenomena may only be ascertainable from experiments or simulations based on the same masses of input data encountered by people.

Moreover, even if a model could solve the same difficult problem that a human does given the same data it would not prove that the model solves the problem in the same way. What to do? Apparently, one necessary test is to require a conjunction of both kinds of evidence—observational or experimental evidence, that learners are exposed to and influenced by a certain set of constraints, and evidence that the same constraints approximate natural human learning and performance when embedded in a simulation model running over a natural body of data. However, in the case of effective but locally weak constraints, the first part of this two-pronged test—experimental or observational demonstration of their human use—might well fail. Such constraints might not be detectable by isolating experiments or in small samples of behavior. Thus, although an experiment or series of observational studies could prove that a particular constraint is used by people, it could not prove that it is not. A useful strategy for such a situation is to look for additional effects predicted by the postulated constraint system in other phenomena exhibited by learners after exposure to large amounts of data.

The Latent Semantic Analysis Model

The model we have used for simulation is a purely mathematical analysis technique. However, we want to interpret the model in a broader and more psychological manner. In doing so, we hope to show that the fundamental features of the theory that we later describe are plausible, to reduce the otherwise magical appearance of its performance, and to suggest a variety of relations to psychological phenomena other than the ones to which we have as yet applied it.

We explicate all of this in a somewhat spiral fashion. First, we try to explain the underlying inductive mechanism of dimensionality optimization upon which the model's power hinges. We then sketch how the model's mathematical machinery operates and how it has been applied to data and prediction. Next, we offer a psychological process interpretation of the model that shows how it maps onto but goes beyond familiar theoretical ideas, empirical principles, findings, and conjectures. We finally return to a more detailed and rigorous presentation of the model and its applications.

An Informal Explanation of the Inductive Value of Dimensionality Optimization

Suppose that Jack and Jill can only communicate by telephone. Jack, sitting high on a hill and looking down at the terrain below estimates the distances separating three houses: A, B, and C. He says that House A is 5 units from both House B and House C, and that Houses B and C are separated by 8 units. Jill uses these estimates to plot the position of the three houses, as shown in the top portion of Figure 1. But then Jack says, "By the way, they are all on the same straight, flat road." Now Jill knows that Jack's estimates must have contained errors and revises her own in a way that uses all three together to improve each one, to 4.5, 4.5, and 9.0, as shown in the bottom portion of Figure 1.

Three distances among three objects are always consistent in

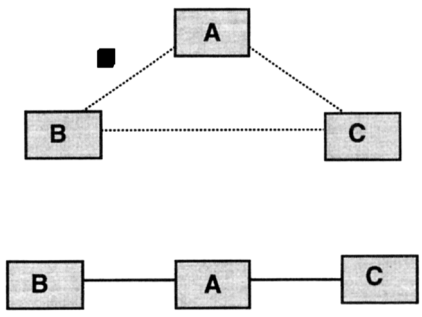

Figure 1. An illustration of the advantage of assuming the correct dimensionality when estimating a set of interpoint distances. Given noisy estimates of AB, AC, and CB, the top portion would be the best guess unless the data source was known to be one-dimensional, in which case the bottom construction would recover the true line lengths more accurately.

two dimensions so long as they obey the triangle inequality (the longest distance must be less than or equal to the sum of the other two). But, knowing that all three distances must be accommodated in one dimension strengthens the constraint (the longest must be exactly equal to the sum of the other two). If the dimensional constraint is not met, the apparent errors in the estimates must be resolved. One compromise is to adjust each distance by the same proportion so as to make two of the lengths add up to the third. The important point is that knowing the dimensionality improves the estimates. Of course, this works the other way around as well. Had the distances been generated from a two- or three-dimensional array (e.g., the road was curved or hilly), accommodating the estimates on a straight line would have distorted their original relations and added error rather than reducing it.

Sometimes researchers have considered dimensionality reduction as a method to reduce computational complexity or for smoothing, that is for simplifying the description of data or interpolating intermediate points (e.g., Church & Hanks, 1990; Grefenstette, 1994; Schütze, 1992a, 1992b). However, as we will see later, choosing the optimum dimensionality, when appropriate, can have a much more dramatic effect than these interpretations would seem to suggest.

Let us now construe the semantic similarity between two words in terms of distance in semantic space: The smaller the distance, the greater the similarity. Suppose we also assume that two words that appear in the same window of discourse—a phrase, a sentence, a paragraph, or what have you—tend to come from nearby locations in semantic space.[1] We could then obtain an initial estimate of the relative similarity of any pair of words by observing the relative frequency of their joint occurrence in such windows.

Given a finite sample of language, such estimates would be quite noisy. Moreover, because of the huge number of words relative to received discourse, many pairwise frequencies would be zero. But two words could also fail to co-occur for a variety of reasons other than thin sampling statistics, with different implications for their semantic similarity. The words might be truly unrelated (e.g., *semantic* and *carburetor*). On the other hand, they might be near-perfect synonyms of which people usually use only one in a given utterance (e.g., *overweight* or *corpulent*), have somewhat different but systematically related meanings (e.g., *purple* and *lavender*), or be relevant to different aspects of the same object (e.g., *gears* and *brakes*) and therefore tend not to occur together (just as only one view of the same object may be present in a given scene). To estimate similarity in this situation, more complex, indirect relations (for example, that both *gears* and *brakes* co-occur with *cars*, but *semantic* and *carburetor* have no common bridge) must somehow be used.

One way of doing this is to take all of the local estimates of distance into account at once. This is exactly analogous to our houses example, and, as in that example, the choice of dimensionality in which to accommodate the pairwise estimates determines how well their mutual constraints combine to give the right results. That is, we suppose that word meanings are represented as points (or vectors; later we use angles rather than distances) in k dimensional space, and we conjecture that it is possible to materially improve estimates of pairwise meaning similarities, and to accurately estimate the similarities among related pairs never observed together, by fitting them simultaneously into a space of the same (k) dimensionality.

This idea is closely related to familiar uses of factor analysis and multi-dimensional scaling, and to unfolding, (J. D. Carroll & Arabie, in press; Coombs, 1964), but using a particular kind of data and writ very large. Charles Osgood (1971) seems to have anticipated such a theoretical development when computational power eventually rose to the task, as it now has. How much improvement results from optimal dimensionality choice depends on empirical issues, the distribution of interword distances, the frequency and composition of their contexts in natural discourse, the detailed structure of distances among words estimated with varying precision, and so forth.

The scheme just outlined would make it possible to build a communication system in which two parties could come to agree on the usage of elementary components (e.g., words, at least up to the relative similarity among pairs of words). The same process would presumably be used to reach agreement on similarities between words and perceptual inputs and between perceptual inputs and each other, but for clarity and simplicity and because the word domain is where we have data and have simulated the process, we concentrate here on word–word relations. Suppose that a communicator possesses a representation of a large number of words as points in a high dimensional space. In generating strings of words, the sender tends to choose words located near each other. Over short time spans, contiguities among output words would reflect closeness in the sender's semantic space. A receiver could make first-order estimates of the distance between pairs by their relative frequency of occurrence in the same temporal contexts (e.g., a paragraph). If the receiver then sets out to represent the results of its statistical knowledge as points in a space of the same or nearly the same dimensionality as that from which it was generated, it may be able to do better, especially, perhaps, in estimating the similarities of words that never or rarely occur together. How much better depends, as we have already said, on matters that can only be settled by observation.

Except for some technical matters, our model works exactly as if the assumption of such a communicative process characterizes natural language (and, possibly, other domains of natural knowledge). In essence, and in detail, it assumes that the psychological similarity between any two words is reflected in the way they co-occur in small subsamples of language, that the source of language samples produces words in a way that ensures a mostly orderly stochastic mapping between semantic similarity and output distance. It then fits all of the pairwise similarities into a common space of high but not unlimited dimensionality. Because, as we see later, the model predicts what words should occur in the same contexts, an organism using such a mechanism could, either by evolution or learning,

[1] For simplicity of exposition, we are intentionally imprecise here in the use of the terms distance and similarity. In the actual modeling, similarity was measured as the cosine of the angle between two vectors in hyperspace. Note that this measure is directly related to the distance between two points described by the projection of the vectors onto the surface of the hypersphere in which they are contained; thus at a qualitative level the two vocabularies for describing the relations are equivalent.

adaptively adjust the number of dimensions on the basis of trial and error. By the same token, not knowing this dimensionality a priori, in our studies we have varied the dimensionality of the simulation model to determine what produces the best results.[2]

More conceptually or cognitively elaborate mechanisms for the representation of meaning also might generate dimensional constraints and might correspond more closely to the mentalistic hypotheses of current linguistic and psycho-linguistic theories. For example, theories that postulate meaningful semantic features could be effectively isomorphic to LSA given the identification of a sufficient number of sufficiently independent features and their accurate quantitative assignment to all the words of a large vocabulary. But suppose that it is not necessary to add such subjective interpretations or elaborations for the model to work. Then LSA could be a direct expression of the fundamental principles on which semantic similarity (as well as other perceptual and memorial relations) are built rather than being a reflection of some other system. It is too early to tell whether the model is merely a mathematical convenience that approximates the effects of true cognitive features and processes or corresponds directly to the actual underlying mechanism of which more qualitative theories now current are themselves but partial approximations. The model we propose is at the computational level described by Marr (1982; see also Anderson, 1990), that is, it specifies the natural problem that must be solved and an abstract computational method for its solution.

A Psychological Description of LSA as a Theory of Learning, Memory, and Knowledge

We give a more complete description of LSA as a mathematical model later when we use it to simulate lexical acquisition. However, an overall outline is necessary to understand a roughly equivalent psychological theory we wish to present first. The input to LSA is a matrix consisting of rows representing unitary event types by columns representing contexts in which instances of the event types appear. One example is a matrix of unique word types by many individual paragraphs in which the words are encountered, where a cell contains the number of times that a particular word type, say *model*, appears in a particular paragraph, say this one. After an initial transformation of the cell entries, this matrix is analyzed by a statistical technique called singular value decomposition (SVD) closely akin to factor analysis, which allows event types and individual contexts to be re-represented as points or vectors in a high dimensional abstract space (Golub, Luk, & Overton, 1981). The final output is a representation from which one can calculate similarity measures between all pairs consisting of either event types or contexts (e.g., word–word, word–paragraph, or paragraph–paragraph similarities).

Psychologically, the data that the model starts with are raw, first-order co-occurrence relations between stimuli and the local contexts or episodes in which they occur. The stimuli or event types may be thought of as unitary chunks of perception or memory. The first-order process by which initial pairwise associations are entered and transformed in LSA resembles classical conditioning in that it depends on contiguity or co-occurrence, but weights the result first nonlinearly with local occurrence frequency, then inversely with a function of the number of different contexts in which the particular component is encountered overall and the extent to which its occurrences are spread evenly over contexts. However, there are possibly important differences in the details as currently implemented; in particular, LSA associations are symmetrical; a context is associated with the individual events it contains by the same cell entry as the events are associated with the context. This would not be a necessary feature of the model; it would be possible to make the initial matrix asymmetrical, with a cell indicating the co-occurrence relation, for example, between a word and closely following words. Indeed, Lund and Burgess (in press; Lund, Burgess, & Atchley, 1995), and Schütze (1992a, 1992b), have explored related models in which such data are the input.

The first step of the LSA analysis is to transform each cell entry from the number of times that a word appeared in a particular context to the log of that frequency. This approximates the standard empirical growth functions of simple learning. The fact that this compressive function begins anew with each context also yields a kind of spacing effect; the association of A and B is greater if both appear in two different contexts than if they each appear twice in one context. In a second transformation, all cell entries for a given word are divided by the entropy for that word, $-\Sigma\, p \log p$ over all its contexts. Roughly speaking, this step accomplishes much the same thing as conditioning rules such as those described by Rescorla & Wagner (1972), in that it makes the primary association better represent the informative relation between the entities rather than the mere fact that they occurred together. Somewhat more formally, the inverse entropy measure estimates the degree to which observing the occurrence of a component specifies what context it is in; the larger the entropy of, say, a word, the less information its observation transmits about the places it has occurred, so the less usage-defined meaning it acquires, and conversely, the less the meaning of a particular context is determined by containing the word.

It is interesting to note that automatic information retrieval methods (including LSA when used for the purpose) are greatly improved by transformations of this general form, the present one usually appearing to be the best (Harman, 1986). It does not seem far-fetched to believe that the necessary transform for good information retrieval, retrieval that brings back text corresponding to what a person has in mind when the person offers one or more query words, corresponds to the functional relations in basic associative processes. Anderson (1990) has drawn attention to the analogy between information retrieval in external systems and those in the human mind. It is not clear which way the relationship goes. Does information retrieval in automatic systems work best when it mimics the circumstances that make people think two things are related, or is there a general logic that tends to make them have similar forms? In automatic information retrieval the logic is usually assumed to be that idealized searchers have in mind exactly the same text as they would like the system to find and draw the words in

[2] Although this exploratory process takes some advantage of chance, there is no reason why any number of dimensions should be much better than any other unless some mechanism like the one proposed is at work. In all cases, the model's remaining parameters were fitted only to its input (training) data and not to the criterion (generalization) test.

their queries from that text (see Bookstein & Swanson, 1974). Then the system's challenge is to estimate the probability that each text in its store is the one that the searcher was thinking about. This characterization, then, comes full circle to the kind of communicative agreement model we outlined above: The sender issues a word chosen to express a meaning he or she has in mind, and the receiver tries to estimate the probability of each of the sender's possible messages.

Gallistel (1990), has argued persuasively for the need to separate local conditioning or associative processes from global representation of knowledge. The LSA model expresses such a separation in a very clear and precise way. The initial matrix after transformation to log frequency divided by entropy represents the product of the local or pairwise processes.[3] The subsequent analysis and dimensionality reduction takes all of the previously acquired local information and turns it into a unified representation of knowledge.

Thus, the first processing step of the model, modulo its associational symmetry, is a rough approximation to conditioning or associative processes. However, the model's next steps, the singular value decomposition and dimensionality optimization, are not contained as such in any extant psychological theory of learning, although something of the kind may be hinted at in some modern discussions of conditioning and, on a smaller scale and differently interpreted, is often implicit and sometimes explicit in many neural net and spreading-activation architectures. This step converts the transformed associative data into a condensed representation. The condensed representation can be seen as achieving several things, although they are at heart the result of only one mechanism. First, the re-representation captures indirect, higher-order associations. That is, if a particular stimulus, X, (e.g., a word) has been associated with some other stimulus, Y, by being frequently found in joint context (i.e., contiguity), and Y is associated with Z, then the condensation can cause X and Z to have similar representations. However, the strength of the indirect XZ association depends on much more than a combination of the strengths of XY and YZ. This is because the relation between X and Z also depends, in a well-specified manner, on the relation of each of the stimuli, X, Y, and Z, to every other entity in the space. In the past, attempts to predict indirect associations by stepwise chaining rules have not been notably successful (see, e.g., Pollio, 1968; Young, 1968). If associations correspond to distances in space, as supposed by LSA, stepwise chaining rules would not be expected to work well; if X is two units from Y and Y is two units from Z, all we know about the distance from X to Z is that it must be between zero and four. But with data about the distances between X, Y, Z, and other points, the estimate of XZ may be greatly improved by also knowing XY and YZ.

An alternative view of LSA's effects is the one given earlier, the induction of a latent higher order similarity structure (thus its name) among representations of a large collection of events. Imagine, for example, that every time a stimulus (e.g., a word) is encountered, the distance between its representation and that of every other stimulus that occurs in close proximity to it is adjusted to be slightly smaller. The adjustment is then allowed to percolate through the whole previously constructed structure of relations, each point pulling on its neighbors until all settle into a compromise configuration (physical objects, weather systems, and Hopfield nets do this too; Hopfield, 1982). It is easy to see that the resulting relation between any two representations depends not only on direct experience with them but with everything else ever experienced. Although the current mathematical implementation of LSA does not work in this incremental way, its effects are much the same. The question, then, is whether such a mechanism, when combined with the statistics of experience, produces a faithful reflection of human knowledge.

Finally, to anticipate what is developed later, the computational scheme used by LSA for combining and condensing local information into a common representation captures multivariate correlational contingencies among all the events about which it has local knowledge. In a mathematically well-defined sense it optimizes the prediction of the presence of all other events from those currently identified in a given context and does so using all relevant information it has experienced.

Having thus cloaked the model in traditional memory and learning vestments, we next reveal it as a bare mathematical formalism.

A Neural Net Analog of LSA

We describe the matrix-mathematics of singular value decomposition used in LSA more fully, but still informally, next and in somewhat greater detail in the Appendix. But first, for those more familiar with neural net models, we offer a rough equivalent in that terminology. Conceptually, the LSA model can be viewed as a simple but rather large three-layered neural net. It has a Layer 1 node for every word type (event type), a Layer 3 node for every text window (context or episode) ever encountered, several hundred Layer 2 nodes—the choice of number is presumed to be important—and complete connectivity between Layers 1 and 2 and between Layers 2 and 3. (Obviously, one could substitute other identifications of the elements and episodes). The network is symmetrical; it can be run in either direction. One finds an optimal number of middle-layer nodes, then maximizes the accuracy (in a least-squares sense) with which activating any Layer 3 node activates the Layer 1 nodes that are its elementary contents, and, simultaneously, vice versa. The conceptual representation of either kind of event, a unitary episode or a word, for example, is a pattern of activation across Layer 2 nodes. All activations and summations are linear.

Note that the vector multiplication needed to generate the middle-layer activations from Layer 3 values is, in general, different from that to generate them from Layer 1 values. Thus a different computation is required to assess the similarity between two episodes, two event types, or an event type and an episode, even though both kinds of entities can be represented as values in the same middle-layer space. Moreover, an event type or a set of event types could also be compared with another of the same or with an episode or combination of episodes by computing their activations on Layer 3. Thus the network can

[3] Strictly speaking, the entropy operation is global, added up over all occurrences of the event type (conditioned stimulus; CS), but it is here represented as a local consequence, as might be the case, for example, if the presentation of a CS on many occasions in the absence of the unconditioned stimulus (US) has its effect by appropriately weakening the local representation of the CS–US connection.

create artificial or "imaginary" episodes, and, by the inverse operations, episodes can generate "utterances" to represent themselves as patterns of event types with appropriately varying strengths. The same things are true in the equivalent singular-value-decomposition matrix model of LSA.

The Singular Value Decomposition (SVD) Realization of LSA

The principal virtues of SVD for this research are that it embodies the kind of inductive mechanisms that we want to explore, that it provides a convenient way to vary dimensionality, and that it can fairly easily be applied to data of the amount and kind that a human learner encounters over many years of experience. Realized as a mathematical data-analysis technique, however, the particular model studied should be considered only one case of a class of potential models that one would eventually wish to explore, a case that uses a very simplified parsing and representation of input and makes use only of linear relations. In possible elaborations one might want to add features that make it more closely resemble what we know or think we know about the basic processes of perception, learning, and memory. It is plausible that complicating the model appropriately might allow it to simulate phenomena to which it has not been applied and to which it currently seems unlikely to give a good account, for example certain aspects of grammar and syntax that involve ordered and hierarchical relations rather than unsigned similarities. However, what is most interesting at this point is how much it does in its present form.

Singular Value Decomposition (SVD)

SVD is the general method for linear decomposition of a matrix into independent principal components of which factor analysis is the special case for square matrices with the same entities as columns and rows. Factor analysis finds a parsimonious representation of all the intercorrelations between a set of variables in terms of a new set of abstract variables, each of which is unrelated to any other but which can be combined to regenerate the original data. SVD does the same thing for an arbitrarily shaped rectangular matrix in which the columns and rows stand for different things, as in the present case one stands for words, the other for contexts in which the words appear. (For those with yet other vocabularies, SVD is a form of eigenvalue–eigenvector analysis or principal components decomposition and, in a more general sense, of two-way, two-mode multidimensional scaling (see J. D. Carroll & Arabie, in press).

To implement the model concretely and simulate human word learning, SVD was used to analyze 4.6 million words of text taken from an electronic version of Grolier's *Academic American Encyclopedia,* a work intended for young students. This encyclopedia has 30,473 articles. From each article we took a sample consisting of (usually) the whole text, or its first 2,000 characters, whichever was less, for a mean text sample length of 151 words, roughly the size of a rather long paragraph. The text data were cast into a matrix of 30,473 columns, each column representing one text sample, by 60,768 rows, each row representing a unique word type that appeared in at least two samples. The cells in the matrix contained the frequency with which a particular word type appeared in a particular text sample. The raw cell entries were first transformed to [ln (1 + cell frequency)/entropy of the word over all contexts]. This matrix was then submitted to SVD and the—for example—300 most important dimensions were retained (those with the highest singular values, i.e., the ones that captured the greatest variance in the original matrix). The reduced dimensionality solution then generates a vector of 300 real values to represent each word and each context. See Figure 2. Similarity was usually measured by the cosine between vectors.[4]

We postulate that the power of the model comes from (optimal) dimensionality reduction. Here is still another, more specific, explanation of how this works. The condensed vector for a word is computed by SVD as a linear combination of data from every cell in the matrix. That is, it is not only the information about the word's own occurrences across documents, as represented in its vector in the original matrix, that determines the 300 values of its condensed vector. Rather, SVD uses everything it can—all linear relations in its assigned dimensionality—to induce word vectors that best predict all and only those text samples in which the word occurs. This expresses a belief that a representation that captures much of how words are used in natural context captures much of what we mean by meaning.

Putting this in yet another way, a change in the value of any cell in the original matrix can, and usually does, change every coefficient in every condensed word vector. Thus, SVD, when the dimensionality is reduced, gives rise to a new representation that partakes of indirect inferential information.

A Brief Note on Neurocognitive and Psychological Plausibility

We, of course, intend no claim that the mind or brain actually computes a SVD on a perfectly remembered event-by-context matrix of its lifetime experience using the mathematical machinery of complex sparse-matrix manipulation algorithms. What we suppose is merely that the mind–brain stores and reprocesses its input in some manner that has approximately the same effect. The situation is akin to the modeling of sensory processing with Fourier decomposition, where no one assumes that the brain uses fast Fourier transform the way a computer does, only that the nervous system is sensitive to and produces a result that reflects the frequency–spectral composition of the input. For

[4] We initially used cosine similarities because they usually work best in the information-retrieval application. Cosines can be interpreted as representing the direction or quality of a meaning rather than its magnitude. For a text segment, that is roughly like what its topic is rather than how much it says about it. For a single word, the interpretation is less obvious. It is worth noting that the cosine measure sums the degree of overlap on each of the dimensions of representation of the two entities being compared. In LSA, the elements of this summation have been assigned equal fixed weights, but it would be a short step to allow differential weights for different dimensions in dynamic comparison operations, with instantaneous weights influenced by, for example, attentional, motivational, or contextual factors. This would bring LSA's similarity computations close to those proposed by Tversky (1977), allowing asymmetric judgments, for example, while preserving its dimension-matching inductive properties.

A

| Word/ | Text sample (context) | | |

knowledge of humanlike word meaning similarities to a significant extent if given a large amount of natural text. Second, supposing it did, would its success depend strongly on the dimensionality of its representation? Third, how would its rate of acquisition compare with that of a human reading the same amount of text? Fourth, how much of its knowledge would come from indirect inferences that combine information across samples rather than directly from the local contextual contiguity information present in the input data?

LSA's Acquisition of Word Knowledge From Text

In answer to the first question, we begin with results from the most successful runs, which used around 300 dimensions, a value that we have often found effective in other applications to large data sets. After training, the model's word knowledge was tested with 80 retired items from the synonym portion of the *Test of English as a Foreign Language (TOEFL)*, kindly provided, along with normative data, by Educational Testing Service (ETS; Landauer & Dumais, 1994, 1996). Each item consists of a stem word, the problem word in testing parlance, and four alternative words from which the test taker is asked to choose that with the most similar meaning to the stem. The model's choices were determined by computing cosines between the vector for the stem word in each item and each of the four alternatives and choosing the word with the largest cosine (except in six cases where the encyclopedia text did not contain the stem, the correct alternative, or both, for which it was given a score of .25). The model got 51.5 correct, or 64.4% (52.5% corrected for guessing by the standard formula [correct-chance/(1-chance)]. By comparison, a large sample of applicants to U.S. colleges from non-English-speaking countries who took tests containing these items averaged 51.6 items correct, or 64.5% (52.7% corrected for guessing). Although we do not know how such a performance would compare, for example, with U.S. school children of a particular age, we have been told that the average score is adequate for admission to many universities. For the average item, LSA's pattern of cosines over incorrect alternatives correlated .44 with the relative frequency of student choices.

Thus, the model closely mimicked the behavior of a group of moderately proficient English readers with respect to judgments of meaning similarity. We know of no other fully automatic application of a knowledge acquisition and representation model, one that does not depend on knowledge being entered by a human but only on its acquisition from the kinds of experience on which a human relies, that has been capable of performing well on a full-scale test used for adults. It is worth noting that LSA achieved this performance using text samples whose initial representation was simply a "bag of words"; that is, all information from word order was ignored, and there was, therefore, no explicit use of grammar or syntax. Because the model could not see or hear, it could also make no use of phonology, morphology, orthography, or real-world perceptual knowledge. More about this later.

The Effect of Dimensionality

The idea underlying our interpretation of the model supposes that the correct choice of dimensionality is important to success.

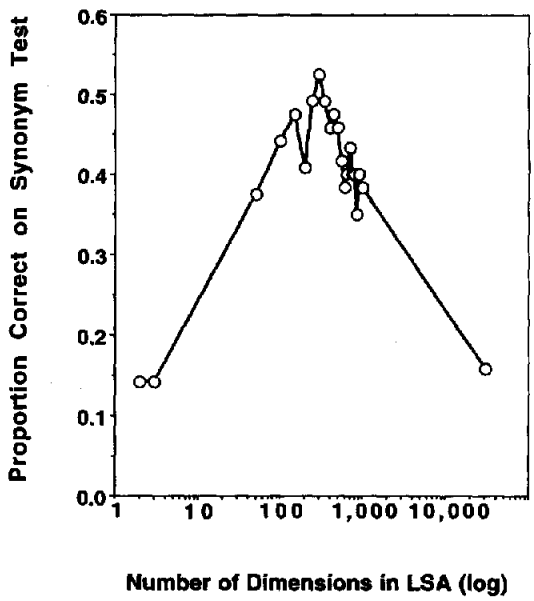

Figure 3. The effect of number of dimensions retained in latent-semantic-analysis (LSA)-singular-value-decomposition (SVD) simulations of word-meaning similarities. The dependent measure is the proportion of 80 multiple-choice synonym test items for which the model chose the correct answer. LSA was trained on text samples from 30,473 articles in an electronic file of text for the *Groliers Academic American Encyclopedia.*

To determine whether it was, the simulation was repeated using a wide range of numbers of dimensions. See Figure 3 (note that the abscissa is on a log scale with points every 50 dimensions in the midregion of special interest). Two or three dimensions, as used, for example in many factor analytic and multidimensional scaling attacks on word meaning (e.g., Deese, 1965; Fillenbaum & Rapoport, 1971; Rapoport & Fillenbaum, 1972) and in the Osgood semantic differential (Osgood, Suci, & Tannenbaum, 1957), resulted in only 13% correct answers when corrected for guessing. More importantly, using too many factors also resulted in very poor performance. With no dimensionality reduction at all, that is, using cosines between rows of the original (but still transformed) matrix, only 16% of the items were correct.[5] Near maximum performance of 45–53%, corrected for guessing, was obtained over a fairly broad region around 300 dimensions. The irregularities in the results (e.g., the dip at 200 dimensions) are unexplained; very small changes in computed cosines can tip LSA's choice of the best test alternative in some cases. Thus choosing the optimal dimensionality of the reconstructed representation approximately tripled the number of words the model learned as compared to using the dimensionality of the raw data.

[5] Given the transform used, this result is similar to what would be obtained by a mutual information analysis, a method for capturing word dependencies often used in computational linguistics (e.g., Church and Hanks, 1990). Because of the transform, this poor result is still better than that obtained by a gross correlation over raw co-occurrence frequencies, a statistic often assumed to be the way statistical extraction of meaning from usage would be accomplished.

Computational constraints prevented assessing points above 1,050 dimensions, except for the full-dimensional case at 30,473 dimensions that could be computed without performing an SVD. However, it is the mid range around the hypothesized optimum dimensionality that is of particular interest here, the matter of determining whether there is a distinct nonmonotonicity in accord with the idea that dimensionality optimization is important. To test the statistical significance of the obvious nonmonotonicity in Figure 3, we fitted separate log functions to the points below and above the observed maximum at 300 dimensions, not including the 300 point itself to avoid the bias of having selected the peak, or the extreme 30,473 point. The positive and negative slopes, respectively, had $r = .98$ ($df = 5$) and $-.86$ ($df = 12$), and associated ps $< .0002$. Thus, it is clear that there is a strong nonmonotonic relation between number of LSA dimensions and accuracy of simulation, with several hundred dimensions needed for maximum performance, but still a small fraction of the dimensionality of the raw data.

The Learning Rate of LSA Versus Humans and Its Reliance on Induction

Next, in order to judge how much of the human learner's problem the model is able to solve, we need to know how rapidly it gains competence relative to human language learners. Even though the model can pass an adult vocabulary test, if it were to require much more data than a human to achieve the same performance one would have to conclude that its induction method was missing something important that humans possess. Unfortunately, we cannot use the ETS normative data directly for this comparison because we don't know how much English their sample of test takers had read, and because, unlike LSA, the ETS students were mostly second-language learners.

For similar reasons, although we have shown that LSA makes use of dimensionality reduction, we do not know how much, quantitatively, this feature would contribute to the problem given the language exposure of a normal human vocabulary learner. We report next some attempts to compare LSA with human word-knowledge acquisition rates and to assess the utility of its inductive powers under normal circumstances.

The rate and sources of schoolchildren's vocabulary acquisition. LSA gains its knowledge of words by exposure to text, a process that is at least partially analogous to reading. How much vocabulary knowledge do humans learn from reading and at what rate? We expand here on the brief summary given earlier. The main parameters of human learning in this major expertise acquisition task have been determined with reasonable accuracy. First note that we are concerned only with knowledge of the relative similarity of individual words taken as units, not with their production or with knowledge of their syntactical or grammatical function, their component spelling, sounds, or morphology or with their real-world pragmatics or referential semantics. That is not to say that these other kinds of word knowledge, which have been the focus of most of the work on lexicon acquisition in early childhood, are unimportant, only that what has been best estimated quantitatively for English vocabulary acquisition as a whole and what LSA has so far been used to simulate is knowledge of the similarity of word meanings.

Reasonable bounds for the long-term overall rate of gain of human vocabulary comprehension, in terms comparable to our LSA results, are fairly well established. The way such numbers usually 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 estimated comprehension vocabularies of young adults, with totals ranging from 40,000 to 100,000 for high school graduates (Nagy & Anderson, 1984; Nagy & Herman, 1987). The variation appears to be largely determined by the size of the dictionaries sampled and to some extent by the way in which words are defined as being separate from each other and by the testing methods employed (see Anglin, 1993; Miller, 1991; and Miller and Wakefield's commentary in Anglin, 1993, for review and critiques). The most common testing methods have been multiple-choice tests much like those of TOEFL, but a few other procedures have been employed with comparable results. Here is one example of an estimation method. Moyer and Landauer (Landauer, 1986) sampled 1,000 words from *Webster's Third New International Dictionary* (1964) and presented them to Stanford University 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 classified correctly 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 of published estimates. The lowest frequently cited estimate is around 40,000 by the last year of high school (Nagy & Anderson, 1984). It appears, however, that all existing estimates are 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)—and most have not adequately counted conventionalized multiword idioms and stock phrases whose meanings cannot or might not be derived from their components.

By simple division, knowing 40,000 to 100,000 words by 20 years of age means adding an average of 7–15 new words a day from age 2 onwards. The rate of acquisition during late elementary and high school years has been estimated at between 3,000 and 5,400 words per year (10–15 per day), with some years in late elementary school showing more rapid gains than the average (Anglin, 1993; Nagy & Herman, 1987; M. Smith, 1941). In summary, it seems safe to assume that, by the usual measures, the total meaning comprehension vocabularies of average fifth-to-eighth–grade students increase by somewhere between 10 and 15 new words per day.

In the LSA simulations every orthographically distinct word, defined as a letter string surrounded by spaces or punctuation marks, is treated as a separate word type. Therefore the most appropriate, although not perfect, correspondence in human word learning is the number of distinct orthographic forms for which the learner must have learned, rather than deduced, the meaning tested by TOEFL. Anglin's (1993; Anglin, Alexander, & Johnson, 1996) recent estimates of schoolchildren's vocabulary attempted to differentiate words whose meaning was stored literally from ones deduced from morphology. This was done by noting when the children mentioned or appeared to use word components during the vocabulary test and measuring their ability to do so when asked. He estimated gains of 9–12

separate learned words per day for first-to-fifth-grade students, without including most proper names or words that have entered the language since around 1980. In addition to the usual factors noted above, there are additional grounds for suspecting that Anglin's estimates may be somewhat low; in particular, the apparent use of morphological analysis could sometimes instead be the result of induced similarity between meanings of independently learned words. For example, LSA computes a relatively high cosine between *independent* and *independence* (cos = .60), *perception* and *perceptual* (cos = .84), *comprehension* and *incomprehensible* (cos = .25; where the average cosine between unrelated words is $\simeq .07 \pm \simeq .04$). LSA, of course has no knowledge of the internal structure of words. Thus children (or adults) asked to tell what *independently* means might think of *independent* not by breaking down *independence* into morphemic components, but because one word reminds them of the other (and adult introspection might fool itself similarly). However, these quibbles are rather beside the point for present purposes. The issue is whether LSA can achieve a rate of learning of word-meaning similarity that approaches or exceeds that of children, and for that purpose the estimates of Anglin, and virtually all others, give an adequate target. To show that its mechanism can do a substantial part of what children accomplish, LSA need only learn a substantial fraction of 10 words per day.

However, a further step in interpreting the LSA–child comparison allows us to more fully resolve the "excess learning" paradox. As mentioned earlier, children in late grade school must acquire most of their new word meanings from reading. The proof is straightforward. The number of different word types in spoken vocabulary is less than a quarter that in the printed vocabulary that people are able to read by the end of high school.[6] Moreover, because the total quantity of heard speech is very large and spoken language undoubtedly provides superior cues for meaning acquisition, such as perceptual correlates, pragmatic context, gestures, and the outright feedback of disambiguating social and tutorial interactions, almost all of the words encountered in spoken language must have been well learned by the middle of primary school. Indeed estimates of children's word understanding knowledge by first grade range upwards toward the tens of thousands used in speech by an average adult (Seashore, 1947). Finally, 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).

It has been estimated that the average fifth-grade child spends about 15 min per day reading in school and another 15 min out of school reading books, magazines, mail, and comic books (Anderson, Wilson, & Fielding, 1988; Taylor, Frye, & Maruyama, 1990). If we assume 30 min per day total for 150 school days and 15 min per day for the rest of the year, we get an average of 21 min per day. At an average reading speed of 165 words per min (Carver, 1990) and a nominal paragraph length of 70 words, they read about 2.5 paragraphs per minute and about 50 per day. Thus, while reading, schoolchildren are adding about one new word to their comprehension vocabulary every 2 min or five paragraphs. 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), one can infer that young readers encounter about one not-yet-known word per paragraph of reading. Thus the opportunity is there to acquire the daily ration. However, this would be an extremely rapid rate of learning. Consider the necessary equivalent list-learning speed. One would have to give children a list of 50 new words, each with one paragraph of exemplary context, and expect them to derive and permanently retain 10–15 sufficiently precise meanings after a single very rapid study trial.

Word meanings are acquired by reading, but how? Several research groups have tried to mimic or enhance the contextual learning of words. The experiments are usually done by selecting nonsense or unknown words at the frontier of grade-level vocabulary knowledge and embedding them in sampled or carefully constructed 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 school books and measured the chance of a word moving from incorrect to correct on a later test as a result of one reading or one hearing (Elley, 1989; Nagy et al., 1985). About one word in 20 paragraphs makes the jump, a rate of 0.05 words per paragraph read. At 50 paragraphs read per day, children would acquire only 2.5 words per day. (Carver and Leibert, 1995, assert that even these rates are high as a result of methodological flaws.)

Thus, experimental attempts intended to produce accelerated vocabulary acquisition have attained less than one half the natural rate, and measurements made under more realistic conditions

[6] From his log-normal model of word frequency distribution and the observations in J. B. Carroll, Davies, and Richmond, (1971), Carroll estimated a total vocabulary of 609,000 words in the universe of text to which students through high school might be exposed. Dahl (1979), whose distribution function agrees with a different but smaller sample of Howes (as cited by Dahl), found 17,871 word types in 1,058,888 tokens of spoken American English, compared to 50,406 in the comparably-sized adult sample of Kucera and Francis (1967). By J. B. Carroll's (1971) model, Dahl's data imply a total of roughly 150,000 word types in spoken English, thus approximately one fourth the total, less to the extent that there are spoken words that do not appear in print. Moreover, the ratio of spoken to printed words to which a particular individual is exposed must be even more lopsided because local, ethnic, favored-TV channels, and family usage undoubtedly restrict the variety of vocabulary more than published works intended for the general school-age readership. If we assume that seventh graders have met a total of 50 million word tokens of spoken English (140 min a day at 100 words per minute for 10 years) then the expected number of occasions on which the they would have heard a spoken word of mean frequency would be about 370. Carroll's estimate for the total vocabulary of seventh-grade texts is 280,000, and we estimate later that typical students would have read about 3.8 million words of print. Thus, the mean number of times they would have seen a printed word to which they might be exposed is only about 14. The rest of the frequency distributions for heard and seen words, although not proportional, would, at every point, show that spoken words have already had much greater opportunity to be learned than printed words, so profit much less from additional experience.

find at best one fourth the normal rate.[7] This leads to the conclusion that much of what the children learned about words from the texts they read must have gone unmeasured in these experiments.

The rate and sources of LSA's vocabulary acquisition. We wish now to make comparisons between the word-knowledge acquisition of LSA and that of children. First, we want to obtain a comparable estimate of LSA's overall rate of vocabulary growth. Second, to evaluate our hypothesis that the model, and by implication, a child, relies strongly on indirect as well as direct learning in this task, we wish to estimate the relative effects of experience with a passage of text on knowledge of the particular words contained in it, and its indirect effects on knowledge of all other words in the language, effects that would not have been measured in the empirical studies of children acquiring vocabulary from text. If LSA learns close to 10 words from the same amount of text that students read, assuming that children use a similar mechanism would resolve the excess-learning paradox.

Because the indirect effects in LSA depend both on the model's computational procedures and on empirical properties of the text it learns from, it is necessary to obtain estimates relevant to a body of text equivalent to what school-age children read. We currently lack a full corpus of representative children's reading on which to perform the SVD. However, we do have access to detailed word-distribution statistics from such a corpus, the one on which the *American Heritage Word Frequency Book* (J. B. Carroll, Davies, & Richman, 1971) was based. By assuming that learners would acquire knowledge about the words in the J. B. Carroll et al. materials in the same way as knowledge about words in the encyclopedia, except with regard to the different words involved, these statistics can provide the desired estimates.

It is clear enough that, for a human, learning about a word's meaning from a textual encounter depends on knowing the meaning of other words. As described above, in principle this dependence is also present in the LSA model. The reduced dimensional vector for a word is a linear combination of information about all other words. Consequently, data solely about other words, for example a text sample containing words Y and Z, but not word X, can change the representation of X because it changes the representations of Y and Z, and all three must be accommodated in the same overall structure. However, estimating the absolute sizes of such indirect effects in words learned per paragraph or per day, and its size relative to the direct effect of including a paragraph actually containing word X calls for additional analysis.

Details of estimating direct and indirect effects. The first step in this analysis was to partition the influences on the knowledge that LSA acquired about a given word into two components, one attributable to the number of passages containing the word itself, the other attributable to the number of passages not containing it. To accomplish this we performed variants of our encyclopedia–TOEFL analysis in which we altered the text data submitted to SVD. We independently varied the number of text samples containing stem words and the number of text samples containing no words from the TOEFL test items. For each stem word from the TOEFL test we randomly selected various numbers of text samples in which it appeared and replaced all occurrences of the stem word in those contexts with a corresponding nonsense word. After analysis we tested the nonsense words by substituting them for the originals in the TOEFL test items. In this way we maintained the natural contextual environment of words while manipulating their frequency. Ideally, we wanted to vary the number of text samples per nonsense word so as to have 2, 4, 8, 16, and 32 occurrences in different repetitions of the experiment. However, because not all stem words had appeared sufficiently often in the corpus, this goal was not attainable, and the actual mean numbers of text samples in the five conditions were 2.0, 3.8, 7.4, 12.8, and 22.2. We also varied the total number of text samples analyzed by the model by taking successively smaller nested random subsamples of the original corpus. We examined total corpus sizes of 2,500; 5,000; 10,000; 15,000; 20,000; 25,000; and 30,473 text samples (the full original corpus). In all cases we retained every text sample that contained any word from any of the TOEFL items.[8] Thus the stem words were always tested by their discriminability from words that had appeared the same, relatively large, number of times in all conditions.

For this analysis we adopted a new, more sensitive outcome measure. Our original figure of merit, the number of TOEFL test items in which the correct alternative had the highest cosine with the stem, mimics human test scores but contains unnecessary binary quantification noise. We substituted a discrimination

[7] Carver and Leibert (1995) have recently put forward a claim that word meanings are not learned from ordinary reading. They report studies in which a standardized 100-item vocabulary test was given before and after a summer program of nonschool book reading. By the LSA model and simulation results to be presented later in this article, one would expect a gain in total vocabulary of about 600 words from the estimated 225,000 words of reading *reported* by their fourth- through sixth-grade participants. Using J. B. Carroll's (1971) model, this would amount to a 0.1%–0.2% gain in total vocabulary. By direct estimates such as Anderson and Freebody (1981), Anglin (1993), Nagy and Anderson (1984), Nagy and Herman (1987), or M. Smith (1941), it would equal about $1/12$ to $1/6$ of a year's increase. Such an amount could not be reliably detected with a 100-item test and 50 students, which would have an expected binomial standard error of around 0.7% or more. Moreover, Carver and Leibert report that the actual reading involved was generally at a relatively easy vocabulary level, which, on a commonsense interpretation, would mean that almost all the words were already known. In terms of LSA, as described later, it would imply that the encountered words were on average at a relatively high point on their learning curves and thus the reading would produce relatively small gains.

[8] Because at least one TOEFL-alternative word occurred in a large portion of the samples, we could not retain all the samples containing them directly, as it would then have been impossible to get small nested samples of the corpus. Instead, we first replaced each TOEFL-alternative word with a corresponding nonsense word so that the alternatives themselves would not be differentially learned, then analyzed the subset corpora in the usual way to obtain vectors for all words. We then computed new average vectors for all relevant samples in the full corpus and finally computed a value for each TOEFL-alternative word other than the stem as the centroid of all the paragraphs in which it appeared in the full corpus. The result is that alternatives other than the stem are always based on the same large set of samples, and the growth of a word's meaning is measured by its progress toward its final meaning, that is, its vector value at the maximum learning point simulated.

ratio measure, computed by subtracting the average cosine between a stem word and the three incorrect alternatives from the cosine between the stem word and the correct alternative, then dividing the result by the standard deviation of cosines between the stem and the incorrect alternatives, that is, (cos stem.correct − mean cos stem.incorrect$_{1-3}$)/(std cos stem.incorrect$_{1-3}$). This yields a z score, which can also be interpreted as a d' measure. The z scores also had additive properties needed for the following analyses.

The results are depicted in Figure 4. Both experimental factors had strong influences; on average the difference between correct and incorrect alternatives grows with both the number of text samples containing the stem word, S, and with additional text samples containing no words on the test, T, and there is a positive interaction between them. For both overall log functions $r > .98$; $F(6)$ for $T = 26.0$, $p << .001$; $F(4)$ for $S = 64.6$, $p << .001$; the interaction was tested as the linear regression of slope on log S as a function of log T, $r^2 = .98$, $F(4) = 143.7$, $p = .001$.) These effects are illustrated in Figure 4 along with logarithmic trend lines for T within each level of S.

Because of the expectable interaction effects—experience with a word helps more when there is experience with other words—quantitative estimates of the total gain from new reading and of the relative contributions of the two factors are only meaningful for a particular combination of the two factors. In other words, to determine how much learning encountering a particular word in a new text sample contributes, one must know how many other text samples with and without that word the learner or model has previously met.

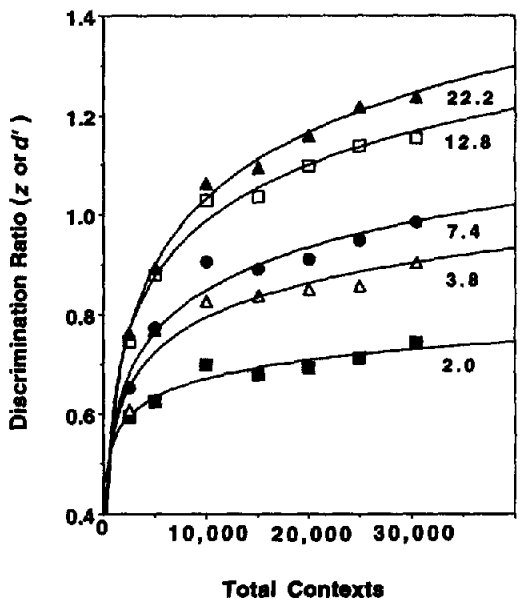

Figure 4. The combined effect in latent semantic analysis (LSA) simulations of the average number of contexts in which a test word appeared (the parameter), and the total number of other contexts, those containing no words from the synonym test items. The dependent measure is the normalized difference in LSA similarity (cosine) of the test words to their correct and incorrect alternatives. The variables were experimentally manipulated by randomly replacing test words with nonsense words and choosing random nested subsamples of total text. The fitted lines are separate empirical log functions for each parameter value.

In the last analysis step, we estimated, for every word in the language, how much the z score for that word increased as a result of including a text sample that contained it and for including a text sample that did not contain it, given a selected point in a simulated schoolchild's vocabulary learning history. We then calculated the number of words that would be correct given a TOEFL-style synonym test of all English words. To anticipate the result, for a simulated seventh grader we concluded that the direct effect of reading a sample on knowledge of words in the sample was an increase of approximately 0.05 words of total vocabulary, and the effect of reading the sample on other words (i.e., all those not in the sample) was a total vocabulary gain of approximately 0.15 words. Multiplying by a nominal 50 samples read, we get a total vocabulary increase of about 10 words per day. Details of this analysis are given next.

Details of LSA simulation of total vocabulary gain. First, we devised an overall empirical model of the joint effects of direct and indirect textual experience that could be fit to the full set of data of Figure 4:

$$z = a(\log b\ T)(\log c\ S) \qquad (1)$$

where T is the total number of text samples analyzed, S is the number of text samples containing the stem word, and a, b, and c are fitted constants (a = 0.128, b = 0.076, c = 31.910 for the present data, least squares fitted by the Microsoft Excel Version 5.0 (1993) iterative equation solver.) Its predictions are correlated with observed z with $r = .98$. To convert its predictions to an estimate of probability correct, we assumed z to be a normal deviate and determined the area under the normal curve to the right of its value minus that of the expected value for the maximum from a sample of three. In other words, we assumed that the cosines for the three incorrect alternatives in each item were drawn from the same normal distribution and that the probability of LSA choosing the right answer is the probability that the cosine of the stem to the correct alternative is greater than the expected maximum of three incorrect alternatives. The overall two-step model is correlated $r = .89$ with observed percentage correct.

Next, we estimated for every word in the language (a) the probability that a word of its frequency appears in the next text sample that a typical seventh grader encounters and (b) the number of times the individual would have encountered that word previously. We then calculated, from Equation 1, (c) the expected increase in z for a word of that frequency as a result of one additional passage containing it and (d) the expected increase in z for a word of that frequency as a result of one additional passage *not* containing it. Finally, we converted z to probability correct, multiplied by the corresponding frequencies, and cumulated gains in number correct over all individual words in the language to get the total vocabulary gains from reading a single text sample.

The J. B. Carroll et al. (1971) data give the frequency of occurrence of each word type in a representative corpus of text read by schoolchildren. Conveniently, this corpus is nearly the same in both overall size, five million words, and in number of word types, 68,000, as our encyclopedia sample (counting, for

the encyclopedia sample, singletons not included in the SVD analysis), so that no correction for sample size, which alters word frequency distributions, was necessary.

Simulating a schoolchild's learning. To simulate the rate of learning for an older grade school child, we assumed that she would have read a total of 3.8 million words, equivalent to 25,000 of our encyclopedia text samples, and set T equal to 25,000 before reading a new paragraph and to 25,001 afterward. We divided the word types in J. B. Carroll et al. (1971) into 37 frequency bands ($< 1, 1, 2, \ldots 20$ and roughly logarithmic thereafter to $> 37{,}000$) and for each band set S equal to an interpolated central frequency of words in the band.[9] We then calculated the expected number of additional words known in each band (the probability correct estimated from the joint-effect model times the probability of occurrence of a token belonging to the band, or the total number of types in the band, respectively) to get (a) the expected direct increase due to one encounter with a test word and (b) the expected increase due to the indirect effect of reading a passage on all other words in the language.[10]

The result was that the estimated direct effect was 0.0007 words gained per word encountered, and the indirect effect was a total vocabulary gain of 0.1500 words per text sample read. Thus the total increase per paragraph read in the number of words the simulated student would get right on a test of all the words in English would be approximately 0.0007×70 (approximate number of words in an average paragraph) + 0.15 = 0.20. Because the average student reads about 50 paragraphs a day (Taylor et al., 1990), the total amounts to about 10 new words per day.

About the accuracy of the simulations. Before further interpreting these results, let us consider their likely precision. The only obvious factors that might lead to overestimated effects are differences between the training samples and text normally read by schoolchildren. First, it is possible that the heterogeneity of the text samples, each of which was drawn from an article on a different topic, might cause a sorting of words by meaning that is more beneficial to LSA word learning than is normal children's text. Counterpoised against this possibility, however, is the reasonable expectation that school reading has been at least partially optimized for children's vocabulary acquisition.

Second, the encyclopedia text samples had a mean of 151 words, and we have equated them with assumed 70 word paragraphs read by schoolchildren. This was done because our hypothesis is that connected passages of text on a particular topic are the effective units of context for learning words and that the best correspondence was between the encyclopedia initial-text samples, usually full short articles, and paragraphs of text read by children. To check the assumption that window-size differences would not materially alter conclusions from the present analysis, we recomputed the TOEFL discrimination ratio results at 300 dimensions for a smaller window size by subdividing the original $\leq$ 2,000 character samples into exhaustive sequential subsets of $\leq$ 500 characters, thus creating a set of 68,527 contexts with a mean of 73 words per sample. The new result was virtually identical to the original value, $z = 0.93$. versus 0.89, corresponding by the models above to about 53% versus 52% correct on TOEFL, respectively.

There are a several reasons to suspect that the estimated LSA learning rate is biased downward rather than upward relative to children's learning. First to continue with the more technical aspects of the analysis, the text samples used were suboptimal in several respects. The crude 2,000 character length cutoff was used because the available machine-readable text had no consistent paragraph or sentence indicators. This resulted in the inclusion of a large number of very short samples, things like "Constantinople: See Istanbul," and of many long segments that contained topical changes that surely would have been signaled by paragraphs in the original.

Of course, we do not know how the human mind chooses the context window. Several alternatives suggest themselves. And it is plausible that the effective contexts are sliding windows rather than the independent samples used here and likely that experienced readers parse text input into phrases, sentences, paragraphs, and other coherent segments rather than arbitrary isolated pieces. Thus, although LSA learning does not appear to be very sensitive to moderate differences in the context window size, window selection was probably not optimized in the reported simulations as well as it is in human reading. The more general question of the effect of window size and manner of selection is of great interest, but requires additional data and analysis.

For the present discussion, more interesting and important differences involve a variety of sources of evidence about word meanings to which human word learners have access but LSA did not. First, of course, humans are exposed to vast quantities of spoken language in addition to printed words. Although we have noted that almost all words heard in speech would be passed on vocabulary tests before seventh grade, the LSA mechanism supposes both that knowledge of these words is still growing slowly in representational quality as a result of new

[9] To estimate the number of words that the learner would see for the very first time in a paragraph, we used the lognormal model proposed by J. B. Carroll (1971) in his introduction to the *Word Frequency Book*. We did not attempt to smooth the other probabilities by the same function because it would have had too little effect to matter, but used a function of the same form to interpolate the center values used to stand for frequency bands.

[10] For example, there are 11,915 word types that appear twice in the corpus. The z for the average word that has appeared twice when 25,000 total samples have been met, according to Equation 1 is 0.75809. If such a word is met in the next sample, which we call a direct effect, it has been met three times, there have been 25,001 total samples, and the word's z increases to 0.83202. By the maximum of three from a normal distribution criterion, its probability of being correct on the TOEFL test rises by 0.029461. But the probability of a given word in a sample being a word of frequency two in the corpus is $(11{,}915 \times 2)/(5 \times 10^6) = 0.0047$, so the direct gain in probability correct for a single word actually encountered attributable to words of frequency two is just 0.000138. However, there is also a very small gain expected for every frequency-two word type that was not encountered, which we call an indirect effect. Adding an additional paragraph makes these words add no occurrences but go from 25,000 to 25,001 samples. By Equation 1, the z for such a word type goes, on average, from 0.75809 to 0.75810, and its estimated probability correct goes up by 7.0674×10^{-6}. But, because there are 11,195 word types of frequency two, the total indirect vocabulary gain is .07912. Finally, we cumulated these effects over all 37 word-frequency bands.

contextual encounters and, more importantly, that new experience with any word improves knowledge of all others.

Second, the LSA analysis treats text segments as mere "bags of words," ignoring all information present in the order of the words, thus making no use of syntax or of the logical, grammatical, discursive, or situational relations it caries. Experts on reading instruction (e.g., Drum & Konopak, 1987; Durkin, 1979) mental abilities (e.g., Sternberg, 1987) and psycholinguistics (e.g., Kintsch & Vipond, 1979; Miller, 1978) have stressed the obvious importance of these factors to the reader's ability to infer word meanings from text. Indeed, Durkin (1983, p. 139) asserts that scrambled sentences would be worthless context for vocabulary instruction (which may well have some validity for human students who have learned some grammar, but clearly is not for LSA).

In the simulations, words were treated as arbitrary units with no internal structure and no perceptual identities; thus LSA could also take no advantage of morphological relations or sound or spelling similarities. Moreover, the data for the simulations was restricted to text, with no evidence provided on which to associate either words or text samples with real-world objects or events or with its own thoughts, emotions, or intended actions as a person might. LSA could make no use of perceptual or experiential relations in the externally referenced world of language or of phonological symbolism (onomatopoeia) to infer the relation between words. Finally, LSA is neither given nor acquires explicitly usable knowledge of grammar (e.g., part-of-speech word classes) or of the pragmatic constraints, such as one-object–one-word, postulated by students of early language acquisition.

Thus, the LSA simulations must have suffered considerable handicaps relative to the modeled seventh-grade student to whom it was compared. Suppose that the seventh grader's extra abilities are used simply to improve the input data represented in Figure 2, for example, by adding an appropriate increment to plurals of words whose singulars appear in a text sample, parsing the input so that verbs and modifiers were tallied jointly only with their objects rather than everything in sight. Such additional information and reduced noise in the input data would improve direct associational effects and presumably be duly amplified by the inductive properties of the dimensionality-matching mechanisms.

Conclusions From the Vocabulary Simulations

There are three important conclusions to be drawn from the results we have described. In descending order of certainty, they are

1. LSA learns a great deal about word meaning similarities from text, an amount that equals what is measured by multiple-choice tests taken by moderately competent English readers.

2. About three quarters of LSA's word knowledge is the result of indirect induction, the effect of exposure to text not containing words used in the tests.

3. Putting all considerations together, it appears safe to conclude that there is enough information present in the language to which human learners are exposed to allow them to acquire the knowledge they exhibit on multiple-choice vocabulary tests. That is, if the human induction system equals LSA in its efficiency of extracting word similarity relations from discourse and has a moderately better system for input parsing and uses some additional evidence from speech and real-world experience, it should have no trouble at all doing the relevant learning it does without recourse to language-specific innate knowledge.

Let us expand a bit on the apparent paradox of schoolchildren increasing their comprehension vocabularies more rapidly than they learn the words in the text they read. This observation could result from either a measurement failure or from induced learning of words not present. The LSA simulation results actually account for the paradox in both ways. First, of course, we have demonstrated very strong inductive learning. But, the descriptive model fitted to the simulation data was also continuous, that is, it assumed that knowledge, in the form of correct placement in the high-dimensional semantic space, is always partial and grows on the basis of small increments distributed over many words. Measurements of children's vocabulary growth from reading have usually looked only at words gotten wrong before reading to see how many of them are gotten right afterwards. In contrast, the LSA simulations computed an increment in probability correct for every word in the potential vocabulary. Thus, it implicitly expresses the hypothesis that word meanings grow continuously and that correct performance on a multiple choice vocabulary test is a stochastic event governed by individual differences in experience, by sampling of alternatives in the test items and by fluctuations, perhaps contextually determined, in momentary knowledge states. As a result, word meanings are constantly in flux, and no word is ever perfectly known. So, for the most extreme example, the simulation computed a probability of one in 500,000 that even the word *the* would be incorrectly answered by some seventh grader on some test at some time.

It is obvious, then, that LSA provides a solution to Plato's problem for at least one case, that of learning word similarities from text. Of course, human knowledge of word meaning is evinced in many other ways, supports many other kinds of performance, and almost certainly reflects knowledge not captured by judgments of similarity. However, it is an open question to what extent LSA, given the right input, can mimic other aspects of lexical knowledge as well.

Generalizing the Domain of LSA

There is no reason to suppose that the mind uses dimensionality optimization only to induce similarities involving words. Many other aspects of cognition would also profit from a means to extract more knowledge from a multitude of local co-occurrence data. Although the full range and details of LSA's implications and applicability await much more research, we give some examples of promising directions, phenomena for which it provides new explanations, interpretations, and predictions. In what follows there are reports of new data, new accounts of established experimental facts, reinterpretation of common observations, and some speculative discussion of how old problems might look less opaque in this new light.

Other Aspects of Lexical Knowledge

By now many readers may wonder how the word similarities learned by LSA relate to meaning. Whereas it is probably impos-

sible to say what word meaning is in a way that satisfies all students of the subject, it is clear that two of its most important aspects are usage and reference. Obviously, the similarity relations between words that are extracted by LSA are based on usage. Indeed, the underlying mathematics can be described as a way to predict the use of words in context, and the only reference of a word that LSA can be considered to have learned in our simulations is reference to other words and to sets of words (although the latter, the contexts of the analysis, may be considered to be coded descriptions of nonlinguistic events). It might be tempting to dismiss LSA's achievements as a sort of statistical mirage, a reflection of the conditions that generate meaning, but not a representation that actually embodies it. We believe that this would be a mistake. Certainly words are most often used to convey information grounded in nonlinguistic events. But to do so, only a small portion of them, and few of the encounters from which the meanings even of those are derived, need ever have been directly experienced in contextual association with the perception of objects, events, or nonlinguistic internal states. Given the strong inductive possibilities inherent in the system of words itself, as the LSA results have shown, the vast majority of referential meaning may well be inferred from experience with words alone. Note that the inductive leaps made by LSA in the simulations were all from purely abstract symbols to other purely abstract symbols. Consider how much more powerful word-based learning would be with the addition of machinery to represent other relations. But for such more elaborate mechanisms to work, language users must agree to use words in the same way, a job much aided by the LSA mechanism.

Even without such extension, however, the LSA model suggests new ways of understanding many familiar properties of language other than word similarity. Here is one homely example. Because, in LSA, word meaning is generated by a statistical process operating over samples of data, it is no surprise that meaning is fluid, that one person's usage and referent for a word is slightly different from the next person's, that one's understanding of a word changes with time, that words drift in both usage and reference over time for the whole community. Indeed, LSA provides a potential technique for measuring the drift in an individual or group's understanding of words as a function of language exposure or interactive history.

Real-World Reference

But still, to be more than an abstract system like mathematics words must touch reality at least occasionally. LSA's inductive mechanism would be valuable here as well. Although not so easily quantified, Plato's problem surely frustrates identification of the perceptual or pragmatic referent of words like *mommy, rabbit, cow, girl, good-bye, chair, run, cry,* and *eat* in the infinite number of real-world situations in which they can potentially appear. What LSA adds to this part of lexicon learning is again its demonstration of the possibility of stronger indirect association than has usually been credited. Because, purely at the word–word level, *rabbit* has been indirectly preestablished to be something like *dog, animal, object, furry, cute, fast, ears,* etc., it is much less mysterious that a few contiguous pairings of the word with scenes including the thing itself can teach the proper correspondences. Indeed, if one judiciously added numerous pictures of scenes with and without rabbits to the context columns in the encyclopedia corpus matrix, and filled in a handful of appropriate cells in the *rabbit* and *hare* word rows, LSA could easily learn that the words *rabbit* and *hare* go with pictures containing rabbits and not to ones without, and so forth. Of course, LSA alone does not solve the visual figure–ground, object parsing, binding, and recognition parts of the problem, but even here it may eventually help by providing a powerful way to generate and represent learned and indirect similarity relations among perceptual features. In any event, the mechanisms of LSA would allow a word to become similar to a perceptual or imaginal experience, thus, perhaps, coming to "stand for" it in thought, to be evoked by it, or to evoke similar images.

Finally, merely using the right word in the right place is, in and of itself, an adaptive ability. A child can usefully learn that the place she lives is Colorado, a college student that operant conditioning is related to learning, a businessperson that TQM is the rage, before needing any clear idea of what these terms stand for. Many well-read adults know that Buddha sat long under a banyan tree (whatever that is) and Tahitian natives lived idyllically on breadfruit and poi (whatever those are). More or less correct usage often precedes referential knowledge (Levy & Nelson, 1994), which itself can remain vague but connotatively useful. Moreover, knowing in what contexts to use a word can function to amplify learning more about it by a bootstrapping operation in which what happens in response provides new context if not explicit verbal correction.

Nonetheless, the implications of LSA for learning pragmatic reference seem most interesting. To take this one step deeper, consider Quine's famous *gavagai* problem. He asks us to imagine a child who sees a scene in which an animal runs by. An adult says "gavagai." What is the child to think *gavagai* means: ears, white, running, or something else in the scene? There are infinite possibilities. In LSA, if two words appear in the same context and every other word in that context appears in many other contexts without them, the two can acquire similarity to each other but not to the rest. This is illustrated in Figures A2 and A4 in the Appendix, which we urge the reader to examine. This solves the part of the problem that is based on Quine's erroneous implicit belief that experiential knowledge must directly reflect first-order contextual associations. What about legs and ears and running versus the whole gavagai? Well, of course, these might actually be what is meant. But by LSA's inductive process, component features of legs, tail, ears, fur, and so forth either before or later are all related to each other, not only because of the occasions on which they occur together, but by indirect result of occasions when they occur with other things and more important, by occasions in which they do not occur at all. Thus the new object in view is not just a collection of unrelated features, each in a slightly different orientation than ever seen before, but a conglomerate of weakly glued features all of which are changed and made yet more similar to each other and to any word selectively used in their presence.

Now consider the peculiar fact that people seem to agree on words for totally private experiences, words like *ache* and *love.* How can someone know that his experience of an ache or of love is like that of his sister? Recognizing that we are having

the same private experience as someone else is an indirect inference, an inference that is often mediated by agreeing on a common name for the experience. We have seen how LSA can lead to agreement on the usage of a word in the absence of any external referent and how it can make a word highly similar to a context even if it never occurs in that context. It does both by resolving the mutual entailments of a multitude of other word–word, word–context, and context–context similarities, in the end defining the word as a point in meaning space that is much the same—but never identical—for different speakers and, perforce, is related to other words and other contextual experiences in much the same way for all. If many times when a mother has a dull pain in her knee, she says "nache," the child may find himself thinking "nache" when having the same experience, even though the mother has never overtly explained herself and never said "nache" when the child's knee hurt. But the verbal and situational contexts of knee pains jointly point to the same place in the child's LSA space as in hers and so does her novel name for the child's similar private experiences. Note, also, how experiences with verbal discourse alone could indirectly influence similarity among perceptual concepts as such, and vice versa, another way to make ears and tails, *aches* and *pains*, run together. Thus, language does not just reflect perception; the two are reciprocally helpful to each other (see D'Andrade, 1993; Lucy & Shweder, 1979, for cogent anthropological evidence on this point).

Conditioning, Perceptual Learning, and Chunking

In this section we take the notion of the model as a homologue of associative learning a few tentative steps further. At this point in the development of the theory, this part must remain conjectural and only roughly specified. The inductive processes of LSA depend on and accrue only to large bodies of naturally interrelated data; thus testing more elaborate and complex models demands more data, computational resources, and time than has been available. Nevertheless, a sketch of some possible implications and extensions shows how the dimensionality-optimizing inductive process might help to explain a variety of important phenomena that appear more puzzling without it and suggests new lines of theory and investigation.

After the dimensionality reduction of LSA every component event is represented as a vector, and so is each context. There is, then, no fundamental difference between components and contexts, except in regard to temporal scale and repeatability; words, for example, are shorter events that happen more than once, and paragraphs are longer events that are almost never met again. Thus, in a larger theoretical framework, or in a real brain, any mental event might serve in either or both roles. For mostly computational reasons, we have so far been able to deal only with two temporal granularities, one nested relation in which repeatability was a property of one type of event and not the other. But there is no reason why much more complex structures, with mental (or neural) events at varying temporal scales and various degrees of repeatability could not exploit the same dimensionality-matching mechanism to produce similarities and generalization among and between psychological entities of many kinds, such as stimuli, responses, percepts, concepts, memories, ideas, images, and thoughts. Because of the mathematical manner in which the model creates representations, a condensed vector representing a context is the same as an appropriately weighted vector average of the condensed vectors of all the events whose local temporal associations constituted it. This has the important property that a new context composed of old units also has a vector representation in (technically, a linear transform of) the space, which in turn gives rise to similarity and generalization effects among new event complexes in an essentially identical fashion to those for two old units or two old contexts. In some examples we give later, the consequences of representing larger segments of experience as a weighted vector sum of the smaller components of which they are built are illustrated. For example, we show how the vector-average representation of a sentence or a paragraph predicts comprehension of a following paragraph, whereas its sharing of explicit words, even when appropriately weighted, does not, and we give examples in which the condensed-vector representation for a whole paragraph determines which of two words it is most similar to, whereas any one word in it may not.

A New Light on Classical Association Theory

Since at least the English associationists, the question of whether association happens by contiguity, similarity, or both has been much argued. LSA provides an interesting answer. In the first instance, similarity is acquired by a process that begins, but only begins, with contiguity. The high-dimensional combination of contiguity data finishes the construction of similarity. But the relations expressed by the high-dimensional representation into which contiguity data are fit are themselves ones of similarity. Thus similarity itself is built of both contiguity and still more similarity. This might explain why an introspectionist, or an experimentalist, could be puzzled about which does what. Even though they are different, the two keep close company, and after sufficient experience, there is a chicken-and-egg relation between their causative effects on representation.

Analogy to Episodic and Semantic Memories

Another interesting aspect of this notion is the light in which it places the distinction between episodic and semantic memory. In our simulations, the model represents knowledge gained from reading as vectors standing for unique paragraph-like samples of text and as vectors standing for individual word types. The word representations are thus semantic, meanings abstracted and averaged from many experiences, while the context representations are episodic, unique combinations of events that occurred only once ever. The retained information about the context paragraph as a single average vector is a representation of gist rather than surface detail. (And, as mentioned earlier, although text passages do not contain all the juice of real biological experience, they are often reasonably good surrogates of nonverbal experience.) Yet both words and episodes are represented by the same defining dimensions, and the relation of each to the other has been retained, if only in the condensed, less detailed form of induced similarity rather than perfect knowledge of history.

Analogy to Explicit and Implicit Memories

In a similar way, the word-versus-context difference might be related to difference between implicit and explicit memories. Retrieving a context vector brings a particular past happening to mind, whereas retrieving a word vector instantiates an abstraction of many happenings irreversibly melded. Thus, for example, recognition that a word came from a particular previously presented list might occur by having the word retrieve one or more context vectors—perhaps experienced as conscious recollections—and evaluating their relation to the word. On the other hand, changes in a word's ability to prime other words occur continuously, and the individual identity of the many occasions that caused the changes, either directly or indirectly, are irretrievable. Although such speculations obviously go far beyond supporting evidence at this point, there is no reason to believe that the processes that rekindle context and word vectors could not be different (indeed, different mathematical operations are required in the SVD model), or even differentially supported by different brain structures. We go no further down this path now than to drop this crumb for future explorations to follow.

Expertise

The theory and simulation results bear interestingly on expertise. Compare the rate of learning a new word, one never encountered before, for a simulated rank novice and an expert reader. Take the rank novice to correspond to the model meeting its second text sample (so as to avoid log 1 in the descriptive model). Assume the expert to have spent 10 years acquiring domain knowledge. Reading 3 hr per day, at 240 words per minute, the expert is now reading his 2,000,001st 70-word paragraph. Extrapolating the model of Equation 1 predicts that the novice gains .14 in probability correct for the new word, the expert .56. Although these extrapolations should not be taken seriously as estimates for human learners because they go outside the range of the empirical data to which the model is known to conform, they nevertheless illustrate the large effects on the ability to acquire new knowledge that can arise from the inductive power inherent in the possession of large bodies of old knowledge. In this case the learning rate, the amount learned about a particular item per exposure to it, is approximately four times as great for the simulated expert as for the simulated novice.

The LSA account of knowledge growth casts a new light on expertise by suggesting that great masses of knowledge contribute to superior performance not only by direct application of the stored knowledge to problem solving, but also by greater ability to add new knowledge to long-term memory, to infer indirect relations among bits of knowledge and to generalize from instances of experience.

Contextual Disambiguation

LSA simulations to date have represented a word as a kind of frequency-weighted average of all its predicted usages. For words that convey only one meaning, this is fine. For words that generate a few closely related meanings, it is a good compromise. This is the case for the vast majority of word types but, unfortunately, not necessarily for a significant proportion of word tokens, because relatively frequent words like *line, fly,* and *bear* often have many senses, as this phenomenon is traditionally described.[11] For words that are seriously ambiguous when standing alone, such as *line,* ones that might be involved in two or more very different meanings with nearly equal frequency, this would appear to be a serious flaw. The average LSA vector for balanced homographs like *bear* can bear little similarity to any of their major meanings. However, we see later that although this raises an issue in need of resolution, it does not prevent LSA from simulating contextual meaning, a potentially important clue in itself.

It seems manifest that skilled readers disambiguate words as they go. The introspective experience resembles that of perceiving an ambiguous figure; only one or another interpretation usually reaches awareness. Lexical priming studies beginning with Ratcliff & McKoon (1978) and Swinney (1979) as well as eye movement studies (Rayner, Pacht, & Duffy, 1994), suggest that ambiguous words first activate multiple interpretations, but very soon settle to that sense most appropriate to their discourse contexts. A contextual disambiguation process can be mimicked using LSA in its current form, but the acquisition and representation of multiple separate meanings of a single word cannot.

Consider the sentence, "The player caught the high fly to left field." On the basis of the encyclopedia-based word space, the vector average of the words in this sentence has a cosine of .37 with *ball,* .31 with *baseball,* and .27 with *hit,* all of which are related to the contextual meaning of *fly,* but none of which is in the sentence. In contrast, the sentence vector has cosines of .17, .18, and .13 with *insect, airplane,* and *bird.* Clearly, if LSA had appropriate separate entries for *fly* that included its baseball sense, distance from the sentence average would choose the right one. However, LSA has only a single vector to represent *fly,* and (as trained on the encyclopedia) it is unlike any of the right words. It has cosines of only .02, .01, and −.02 respectively with *ball, baseball,* and *hit* (compared to .69, .53 and .24, respectively with *insect, airplane,* and *bird*). The sentence representation has correctly caught the drift, but the single averaged-vector representation for the word *fly,* which falls close to midway between *airplane* and *insect,* is nearly orthogonal to any of the other words. More extensive simulations of LSA-based contextual disambiguation and their correlations with empirical data on text comprehension are described later. Meanwhile, we sketch several ways in which LSA might account for multiple meanings of the same word: first a way in which it might be extended to induce more than one vector for a word, then ways in which a single vector as currently computed might give rise to multiple meanings.

It is well-known that, for a human reader, word senses are almost always reliably disambiguated by local context. Usually one or two words to either side of an ambiguous word are enough to settle the overall meaning of a phrase (Choueka & Lusignan, 1985). Context-based techniques for lexical disam-

[11] For example, among the most frequent 400 words in the Kucera and Francis (1967) count, at least 60 have two or more common meanings, whereas in a sample of 400 that appeared only once in the corpus there were no more than 10.

biguation have been tried in computational linguistic experiments with reasonably good results (e.g., Grefenstette, 1994; Schütze, 1992a; Schütze & Pedersen, 1995; Walker & Amsler, 1986). However, no practical means for automatically extracting and representing all the different senses of all the words in a language from language experience alone has emerged.

How might separate senses be captured by an LSA-like model? Suppose that the input for LSA were a three-way rather than a two-way matrix, with columns of paragraphs, ranks of all the phrases that make up all the paragraphs, and rows of all the word types that make up all the phrases. Partway between paragraphs and words, phrases would seldom, but sometimes, repeat. Cells would contain the number of times that a word type appeared in a particular phrase in a particular paragraph. (A neural network equivalent might have an additional layer of nodes. Note that in either case, the number of such intermediate vectors would be enormous, a presently insurmountable computational barrier.)

The reduced-dimensionality representation would constitute a predictive device that would estimate the likelihood of any word occurring in any phrase context or any paragraph, or any phrase occurring in any paragraph, whether they had occurred there in the first place or not. The idea is that the phrase-level vectors would carry distinctions corresponding approximately to differential word senses. In simulating text comprehension, a dynamic performance model might start with the average of the words in a paragraph and, using some constraint satisfaction method, arrive at a representation of the paragraph as a set of imputed phrase vectors and their average.

A very different, much simpler, possibility is that each word has but a single representation, but because LSA representations have very high dimensionality, the combination of a word with a context can have very different effects on the meaning of different passages. Consider the sentences, "The mitochondria are in the cells," versus "The monks are in the cells," in which abstract semantic dimensions of the context determine the sense of *cells* as biological or artificial objects. In one case the overall passage-meaning vector has a direction intermediate between that of *mitochondria* and that of *cells,* in the other case between *monks* and *cells.* If *mitochondria* and *monks* are in orthogonal planes in semantic space, the resultant vectors are quite different. Now suppose that the current context-specific meaning of *cells*—and perhaps its conscious expression—is represented by the projection of its vector onto the vector for the whole passage; that is, only components of meaning that it shares with the context, after averaging, comprise its disambiguated meaning. In this way, two or more distinct senses could arise from a single representation, the number and distinctions among senses depending only on the variety and distinctiveness of different contexts in which the word is found. In this interpretation, the multiple senses described by lexicographers are categorizations imposed on the contextual environments in which a word is found.

Put another way, a 300-dimensional vector has plenty of room to represent a single orthographic string in more than one way so long as context is sufficient to select the relevant portion of the vector to be expressed. In addition, it might be supposed that the relations among the words in a current topical context would be subjected to a local re-representation process, a secondary SVD-like condensation, or some other mutual constraint satisfaction process using the global cosines as input that would have more profound meaning-revision effects than simple projection.

Finally, the contextual environment of a word might serve to retrieve related episode representations that would, by the same kinds of processes, cause the resultant meaning, and perhaps the resultant experience, to express the essence of a particular subset of past experiences. Given an isolated word, the system might settle competitively on a retrieved vector for just one or the average of a concentrated cluster of related episodes, thus giving rise to the same phenomenology, perhaps by the same mechanism, as the capture quality of ambiguous visual figures. Thus the word *cell* might give rise to an image of either a microscopic capsule or a room.

A resolution of which, if any, of these hypothetical mechanisms accounts for multiple word-meaning phenomena is beyond the current state of LSA theory and data; the moral of the discussion is just that LSA's single-vector representation of a word is not necessarily a fatal or permanent flaw. Whereas some of the evidence to follow inclines us to the single-representation view, we consider the issue as distinctly open.

Text Comprehension: An LSA Interpretation of Construction-Integration Theory

Some research has been done using LSA to represent the meaning of segments of text larger than words and to simulate behaviors that might otherwise fall prey to the ambiguity problem. In this work, individual word senses are not separately identified or represented, but the overall meaning of phrases, sentences, or paragraphs is constructed from a linear combination of their words. By hypothesis, the various unintended-meaning components of the many different words in a passage tend to be unrelated and point in many directions in meaning hyperspace, whereas their vector average reflects the overall topic or meaning of the passage. We recount two studies illustrating this strategy. Both involve phenomena that have previously been addressed by the construction-integration (CI) model (Kintsch, 1988). In both, the current version of LSA, absent any mechanism for multiple-word-sense representation, is used in place of the intellectually coded propositional analyses of CI.

Predicting coherence and comprehensibility. Foltz, Kintsch, and Landauer, in an unpublished study (1993), reanalyzed data from experiments on text comprehension as a function of discourse coherence. As part of earlier studies (McNamara, Kintsch, Butler-Songer, & Kintsch, 1996), a single short text about heart function had been reconstructed in four versions that differed greatly in coherence according to the propositional analysis measures developed by Van Dijk and Kintsch (1983). In coherent passages, succeeding sentences used concepts introduced in preceding sentences so that the understanding of each sentence and of the overall text—the building of the text base and situation model in CI terms—could proceed in a gradual, stepwise fashion. In less coherent passages, more new concepts were introduced without precedent in the propositions of preceding sentences. The degree of coherence was assessed by the number of overlapping concepts in propositions of successive sentences. Empirical comprehension tests with college student

readers established that the relative comprehensibility of the four passages was correctly ordered by their propositionally estimated coherence.

In the reanalysis, sentences from a subcorpus of 27 encyclopedia articles related to the heart were first subjected to SVD and a 100-dimensional solution used to represent the contained words. Then each sentence in the four experimental paragraphs was represented as the average of the vectors of the words it contained. Finally, the coherence of each paragraph was reestimated as the average cosine between its successive sentences. Figure 5 shows the relation of this new measure of coherence to the average empirical comprehension scores for the paragraphs. The LSA coherence measure corresponds well to measured comprehensibility. In contrast, an attempt to predict comprehensibility by correlating surface-structure word types in common between successive sentences (i.e., computing cosines between vectors in the full-dimension transformed matrix), also shown in Figure 5, fails, largely because there is little overlap at the word level. LSA, by capturing the central meaning of the passages appears to reflect the differential relations among sentences that led to comprehension differences.

Simulating contextual word disambiguation and sentential meaning inference. Another reanalysis illustrates this reinterpretation of CI in LSA terms more directly with a different data set. Till, Mross, and Kintsch (1988) performed semantic priming experiments in which readers were presented word by word with short paragraphs and interrupted at strategically placed points to make lexical decisions about words related either to one or another of two senses of a just-presented homographic word or to words not contained in the passages but related inferentially to the story situation that a reader would presumably assemble in comprehending the discourse up to that point. They also varied the interval between the last text word shown and the target for lexical decision. Here is an example of two matched text paragraphs and the four target words for lexical decisions used in conjunction with them.

1. The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the *mole.*

2. The patient sensed that this was not a routine visit. The doctor hinted that there was serious reason to remove the *mole.*

Targets for lexical decision: *ground, face; drown, cancer*

Across materials, Till et al. (1988) balanced the materials by switching words and paragraphs with different meanings and included equal numbers of nonwords. In three experiments of this kind, the principal findings were (a) in agreement with Ratcliff and McKoon (1978) and Swinney (1979), words related to both senses of an ambiguous word were primed immediately after presentation, (b) after about 300 ms only the context appropriate associates remained significantly primed, and (c) words related to inferred situational themes were not primed at short intervals, but were at delays of 1 s.

The standard CI interpretation of these results is that in the first stage of comprehending a passage—construction—multiple nodes representing all senses of each word are activated in long-term memory, and in the next stage—integration—iterative excitation and inhibition among the nodes leads to dominance of appropriate word meanings and finally to creation of a propositional structure representing the situation described by the passage.

LSA as currently developed is, of course, mute on the temporal dynamics of comprehension, but it does provide an objective way to represent, simulate, and assess the degree of semantic similarity between words and between words and longer passages. To illustrate, an LSA version of the CI account for the Till et al. (1988) experiment might go like this:

1. First, a central meaning for each graphemic word type is retrieved: the customary vector for each word. Following this, there are two possibilities, depending on whether one assumes single or multiple representations for words.

2. Assuming only a single, average representation for each word, the next step is computation of the vector average for all words in the passage. As this happens, words related to the average meanings being generated, including both appropriate relatives of the homograph and overall "inference" words, become activated, while unrelated meanings, including unrelated associates of the homograph, decline.

On the other interpretation, an additional stage is inserted between these two in which the average meaning for some or all of the words in the passage disambiguates the separate words individually, choosing a set of senses that are then combined. The stimulus asynchrony data of Till et al. (1988) seems to suggest the latter interpretation in that inappropriate homograph relatives lose priming faster than inference words acquire it, but there are other possible explanations for this result, in particular that the overall passage meaning simply evolves slowly with the most holistic interpretations emerging last. In any event, the current LSA representation can only simulate the meaning relations between the words and passages and is indifferent to which

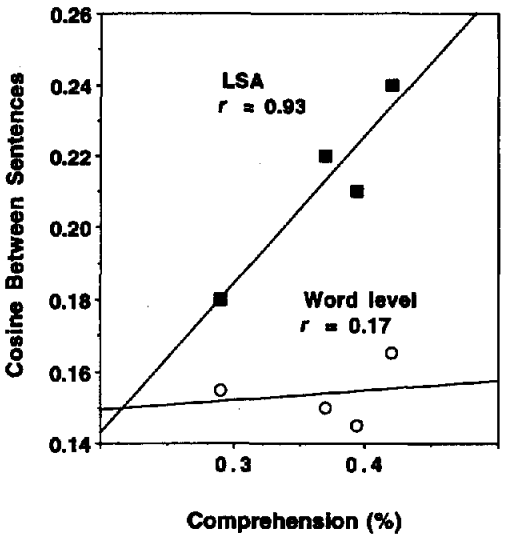

Figure 5. Prediction of measured text comprehensibility of a set of experimentally altered text passages taken from McNamara et al. (1996). Predictions were based on the similarity of each sentence to that of the succeeding sentence, putative measures of conceptual coherence. For latent semantic analysis (LSA), sentences were represented by the average of the LSA-derived vectors of the words they contained. The control condition (word level) used the same analysis but without dimension reduction.

of these alternatives, or some other, is involved in the dynamics of comprehension.

In either case, LSA predicts that (a) there should be larger cosines between the homographic word and both of its related words than between it and control words, (b) the vector average of the passage words coming before the homographic word should have a higher cosine with the context-relevant word related to it than to the context-irrelevant word, and (c) the vector average of the words in a passage should have a higher cosine with the word related to the passage's inferred situational meaning than to control words.

These predictions were tested by computing cosines based on word vectors derived from the encyclopedia analysis and comparing the differences in mean similarities corresponding to the word–word and passage–word conditions in Till et al. (1988, Experiment 1). There were 28 pairs of passages and 112 target words. For the reported analyses, noncontent words such as *it, of, and, to, is, him,* and *had* were first removed from the passages, then vectors for the full passages up to or through the critical homograph were computed as the vector average of the words. The results are shown in Table 1. Here is a summary.

1. Average cosines between ambiguous homographs and the two words related to them were significantly higher than between the homographs and unrelated words (target words for other sentence pairs). The effect size for this comparison was at least as large as that for priming in the Till et al. (1988) experiment.

2. Homograph-related words that were also related to the meaning of the paragraph had significantly higher cosines with the vector average of the passage than did paired words related to a different sense of the homograph. For 37 of the 56 passages the context-appropriate sense related word had a higher cosine with the passage preceding the homograph than did the inappropriate sense-related word ($p = .01$). (Note that these are relations to particular words, such as *face,* that are used to stand—imperfectly at best—for the correct meaning of *mole,* rather than the hypothetical correct meaning itself. Thus, for all we know, the true correct disambiguation, as a point in LSA meaning space, was always computed).

3. To assess the relation between the passages and the words ostensibly related to them by situational inference, we computed cosines between passage vector averages and the respective appropriate and inappropriate inference target words and between the passages and unrelated control words from passages displaced by two in the Till et al. (1988) list. On average, the passages were significantly closer to the appropriate than to either the inappropriate inferentially related words or unrelated control words (earlier comment relevant here as well).

These word and passage relations are fully consistent with either LSA counterpart of the construction-integration theory as outlined above. In particular, they show that an LSA based on (only) 4.6 million words of text produced representations of word meanings that would allow the model to mimic human performance in the Till et al. (1988) experiment given the right activation and interaction dynamics. Because homographs are similar to both tested words presumably related to different meanings, they presumably could activate both senses. Because the differential senses of the homographs represented by their related words are more closely related to the average of words in the passage from which they came, the LSA representation of the passages would provide the information needed to select the homograph's contextually appropriate associate. Finally, the LSA representations of the average meaning of the passages are similar to words related to meanings thought to be inferred from mental processing of the textual discourse. Therefore, the LSA representation of the passages must also be related to the overall inferred meaning.

Some additional support is lent to these interpretations by findings of Lund, Burgess, and colleagues (Lund & Burgess, in press; Lund et al., 1995) who have mimicked other priming

Table 1
LSA Simulation of Till et al. (1988) Sentence and Homograph Priming Experiment

	Sense targets		Inference targets		
Prime	Right (A)	Wrong (B)	Right (C)	Wrong (D)	Unrelated (control)
Homograph alone	.20	.21	.09	.05	.07 p vs. A or B < .00001 z = .89
Full passage with homograph	.24	.21	.21	.14 p vs. C = .0008 z = 1.59	.15 p vs. C = .0005 z = .55
Full passage without homograph	.21	.15 p vs. A = .006 z = .48	.21	.14 p vs. C = .0002 z = .69	.16 p vs. C = .002 z = .46

Note. Simulated discourse was from Till, Kintsch, and Mross (1988). Cell entries are latent semantic analysis (LSA) cosines between words, or words and sentences, based on a large text-corpus analysis. Targets in Columns A and B were common associates of the homographic word ending the sentence, either related or not to the sense of the passage. Targets in Columns C and D were words not in a sentence but intuitively related, or not, to its overall inferred meaning. Probabilities are based on individual two-sample, one-tailed t-tests, dfs $\geq$ 54. Differences < .05 and without stated p values had $p > .09$.

data using a high-dimensional semantic model, HAL, that is related to LSA.[12] Lund et al. derived 200 element vectors to represent words from analysis of 160 million words from Usenet newsgroups. They first formed a word–word matrix from a 10-word sliding window in which the co-occurrence of each pair of words was weighted inversely with the number of intervening words. They reduced the resulting 70,000-by-70,000 matrix to one of 70,000 by 200 simply by selecting only the 200 columns (following words) with the highest variance. In a series of simulations and experiments, they have been able to mimic semantic priming results that contrast pairs derived from free-association norms and pairs with intuitively similar meanings, interpreting their high-dimensional word vectors as representing primarily (judged) semantic relatedness.

At least two readings of the successful mimicking of lexical priming relations by high-dimensional, semantic-space similarities are possible. One is that some previous findings on textual word and discourse processing may have been a result of word-to-word and word-set-to-word similarities rather than the more elaborate cognitive-linguistic processes of syntactic parsing and sentential semantic meaning construction that have usually been invoked to explain them. Word and, especially, word-set semantic relations were not conveniently measurable prior to LSA and could easily have been overlooked. However, we believe it would be incorrect to suggest that previous text-processing results are in any important sense artifactual. For one thing, even the more cognitively elaborate theories, such as CI, depend on semantic relations among words, which are customarily introduced into the models on the basis of expert subjective judgments or human association norms. LSA might be viewed as providing such models with a new tool for more objective simulation, for acquiring word–word relations from input data like that used by humans rather than "black-box" outputs of some of the processes we wish to understand. For another, we have no intention of denying an important role to syntax-using, meaning-construction processes. We are far from ready to conclude that LSA's representation of a passage as a weighted vector average of the words in it is a complete model of a human's representation of the same passage.

On the other hand, we think it would be prudent for researchers to attempt to assess the degree to which language-processing results can be attributed to word and word-set meaning relations and to integrate these relations into accounts of psycholinguistic phenomena. We also believe that extensions of LSA, including extensions involving iterative construction of context-dependent superstructures, and dynamic processes for comprehension, might in many cases present a viable alternative to psycholinguistic models based on more traditional linguistic processes and representations.

Mimicking the representation of single-digit Arabic numerals. The results described up to here have assessed the LSA representation of words primarily with respect to the similarity between two words or between a word and the combination of a set of words. But a question still needs asking as to the extent to which an LSA representation corresponds to all or which aspects of what is commonly understood as a word's meaning. The initial performance of the LSA simulation on TOEFL questions was as good as that of students who were asked to judge similarity of meaning. This suggests that the students did not possess more or better representations of meaning for the words involved, that the LSA representation exhausted the usable meaning for the judgment. However, the students had limited abilities and the tests had limited resolution and scope; thus much of each word's meaning may have gone undetected on both sides. The rest of the simulations, for example the predictions of paragraph comprehension and sentence-inference priming, because they also closely mimic human performances usually thought to engage and use meaning, add weight to the hypothesis that LSA's representation captures a large component of human meaning. Nevertheless, it is obvious that the issue is far from resolved.

At this point, we do no more than to add one more intriguing finding that demonstrates LSA's representation of humanlike meaning in a rather different manner. Moyer & Landauer (1967) reported experiments in which participants were timed as they made button presses to indicate which of two single-digit numerals was the larger. The greater the numerical difference between the two, the faster was the average response. An overall function that assumed that single-digit numerals are mentally represented as the log of their arithmetic values and judged as if they were line lengths fit the data nicely. But why should people represent digits as the logs of their numerical value? It makes no apparent sense either in terms of the formal properties of mathematics, of what people have learned about these symbols for doing arithmetic, or for their day-to-day role in counting or communication of magnitudes.

A model of meaning acquisition and generation should be able to account for nonobvious and apparently maladaptive cases as well as those that are intuitively expectable. What relations among the single-digit number symbols does LSA extract from text? To find out, we performed a multidimensional scaling on a matrix of all 36 dissimilarities (defined as 1-LSA cosine) between the digits 1 through 9 as encountered as single isolated characters in the encyclopedia text sample. A three-dimensional solution accounted for almost all the interdigit dissimilarities (i.e., their local structure, not the location or orientation of that structure in the overall space). Projections of the nine digit representations onto the first (strongest) dimension of the local structure are shown in Figure 6.

Note first that the digits are aligned in numerical order on this dimension, second that their magnitudes on the dimension are nearly proportional to the log of their numerical values. Clearly, the LSA representation captures the connotative meaning reflected in inequality judgment times. The implication is that the reason that people treat these abstract symbols as having continuous analog values on a log scale is simply that the statistical properties of their contextual occurrences implies these relations. Of course, this raises new questions, in particular, where or how generated is the memory representation that allows people to use numerals to add and subtract with digital accuracy:

[12] There is a direct line of descent between LSA and the HAL model of Burgess and colleagues (Lund & Burgess, in press; Lund et al., 1995). They credit an unpublished article of H. Schütze as the inspiration for their method of deriving semantic distance from large corpora, and Schütze, in the same and other articles (e.g., 1992a), cites Deerwester et al. (1990), the initial presentation of the LSA method for information retrieval.

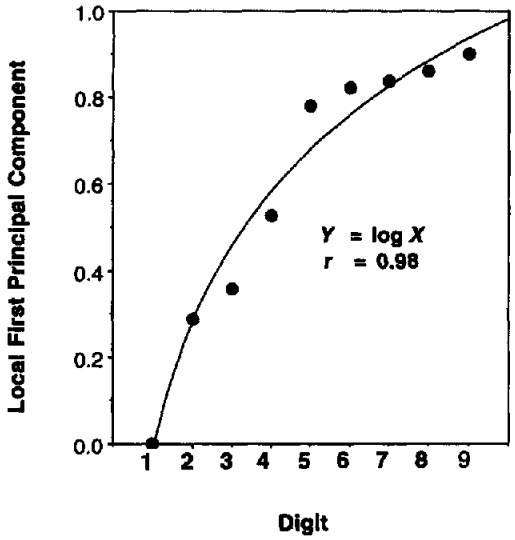

Figure 6. The dissimilarities (1-cosine) between all pairs of latent semantic analysis (LSA) vectors representing the single-digit numerals 1–9, as derived from large text-corpus training, were subjected to multidimensional scaling. The projection of the point for each numeral onto the first principal component of this LSA subspace is shown. (The scale of the dimension has been linearly adjusted to an arbitrary 0–1 range.) The numeral representations align in numerical order and scale as their logs, reflecting, it is proposed, the dimension of meaning tapped by inequality judgment times as observed by Moyer and Landauer (1967).

in another projection, in the representation of number-fact phrases, or somewhere or somehow else?

It must be noted that the frequency of occurrence in English of the Arabic numerals 1–9 is also related to the log of their numerical value, larger numbers having smaller frequencies (Davies, 1971), in which case it might appear that people's judgment of numeral differences are in reality judgments that the one with the smaller frequency is the larger. However, this possibility does not greatly affect the point being made here, which is that a particular context-conditioned projection of the LSA representations revealed a component dimension related to a meaning-based performance, judgment of relative size, that goes beyond judgment of the pairwise similarities of the objects.

A hint for future research that we take from this result is that there may often be projections of word meanings onto locally defined dimensions that create what from other perspectives may be puzzling combinations of meaning. For example, the reading of a lexically ambiguous word in a sentence or the effect of an otherwise anomalous word in a metaphorical expression might depend, not on the position of the word in all 300 dimensions, but on its position in a perhaps temporary local subspace that best describes the current context. This conjecture awaits further pursuit.

Summary

We began by describing the problem of induction in knowledge acquisition, the fact that people appear to know much more than they could have learned from temporally local experiences. We posed the problem concretely with respect to the learning of vocabulary by school-age children, a domain in which the excess of knowledge over apparent opportunity to learn is quantifiable and for which a good approximation to the total relevant experience available to the learner is also available to the researcher. We then proposed a new basis for long-range induction over large knowledge sets containing only weak and local constraints at input. The proposed induction method depends on reconstruction of a system of multiple similarity relations in a high dimensional space. It is supposed that the co-occurrence of events, words in particular, in local contexts is generated by and reflects their similarity in some high-dimensional source space. By reconciling all the available data from local co-occurrence as similarities in a space of nearly the same dimensionality as the source, a receiver can, we propose, greatly improve its estimation of the source similarities over their first-order estimation from local co-occurrence. The actual value of such an induction and representational scheme is an empirical question and depends on the statistical structure of large natural bodies of information. We hypothesized that the similarity of topical or referential meaning ("aboutness") of words is a domain of knowledge in which there are very many indirect relations among a very large number of elements and, therefore, one in which such an induction method might play an important role.

We implemented the dimensionality-optimizing induction method as a mathematical matrix-decomposition method called singular value decomposition (SVD) and tested it by simulating the acquisition of vocabulary knowledge from a large body of text. After analyzing and re-representing the local associations between some 60,000 words and some 30,000 text passages containing them, the model's knowledge was assessed by a standardized synonym test. The model scored as well as the average of a large sample of foreign students who had taken this test for admission to U.S. colleges. The model's synonym test performance depended strongly on the dimensionality of the representational space into which it fit the words. It did very poorly when it relied only on local co-occurrence (too many dimensions), well when it assumed around 300 dimensions, and very poorly again when it tried to represent all its word knowledge in much less than 100 dimensions. From this, we concluded that dimensionality-optimization can greatly improve the extraction and representation of knowledge in at least one domain of human learning.

To further quantify the model's (and thus the induction method's) performance, we simulated the acquisition of vocabulary knowledge by school-children. The model simulations learned at a rate—in total vocabulary words added per paragraph read—approximating that of children and considerably exceeding learning rates that have been attained in laboratory attempts to teach children word meanings by context. Additional simulations showed that the model, when emulating a late-grade school child, acquired most of its knowledge about the average word in its lexicon through induction from data about other words. One evidence of this was an experiment in which we varied the number of text passages either containing or not containing tested words and estimated that three fourths of total vocabulary gain from reading a passage was in words not in the paragraph at all.

Given that the input to the model was data only on occurrence

of words in passages, so that LSA had no access to word-similarity information based on spoken language, morphology, syntax, logic, or perceptual world knowledge, all of which can reasonably be assumed to be additional evidence that a dimensionality-optimizing system could use, we conclude that this induction method is sufficiently strong to account for Plato's paradox—the deficiency of local experience—at least in the domain of knowledge measured by synonym tests.

Based on this conclusion, we suggested an underlying associative learning theory of a more traditional psychological sort that might correspond to the mathematical model and offered a sample of conjectures as to how the theory would generate novel accounts for aspects of interesting psychological problems, in particular for language phenomena, expertise, and text comprehension. Then, we reported some reanalyses of human text processing data in which we illustrated how the word and passage representations of meaning derived by LSA can be used to predict such phenomena as textual coherence and comprehensibility and to simulate the contextual disambiguation of homographs and generation of the inferred central meaning of a paragraph. Finally, we showed how the LSA representation of digits can explain why people apparently respond to the log of digit values when making inequality judgments.

At this juncture, we believe the dimensionality-optimizing method offers a promising solution to the ancient puzzle of human knowledge induction. It still remains to determine how wide its scope is among human learning and cognition phenomena: Is it just applicable to vocabulary, or to much more, or, perhaps, to all knowledge acquisition and representation? We would suggest that applications to problems in conditioning, association, pattern and object recognition, contextual disambiguation, metaphor, concepts and categorization, reminding, case-based reasoning, probability and similarity judgment, and complex stimulus generalization are among the set where this kind of induction might provide new solutions. It still remains to understand how a mind or brain could or would perform operations equivalent in effect to the linear matrix decomposition of SVD and how it would choose the optimal dimensionality for its representations, whether by biology or an adaptive computational process. And it remains to explore whether there are better modeling approaches and input representations than the linear decomposition methods we applied to unordered bag-of-words inputs. Conceivably, for example, different input and different analyses might allow a model based on the same underlying induction method to derive aspects of grammar and syntactically based knowledge. Moreover, the model's objective technique for deriving representations of words (and perhaps other objects) offers attractive avenues for developing new versions and implementations of dynamic models of comprehension, learning, and performance. On the basis of the empirical results and conceptual insights that the theory has already provided, we believe that such explorations are worth pursuing.

References

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

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Huston (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*(3), 285–303.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*(10, Serial No. 238).

Anglin, J. M., Alexander, T. M., & Johnson, C. J. (1996). *Word learning and the growth of potentially knowable vocabulary.* Unpublished manuscript.

Angluin, D., & Smith, C. H. (1983). Inductive inference: Theory and methods. *Computing Surveys, 15,* 237–269.

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

Bookstein, A., & Swanson, D. R. (1974). Probabilistic models for automatic indexing. *Journal of the American Association for Information Science, 25,* 312–318.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carroll, J. B. (1971). Statistical analysis of the corpus. In J. B. Carroll, P. Davies, & B. Richman (Eds.), *Word frequency book* (pp. xxii–xl). New York: Houghton Mifflin and American Heritage.

Carroll, J. B., Davies, P., & Richman, B. (Eds.). (1971). *Word frequency book.* New York: Houghton Mifflin and American Heritage.

Carroll, J. D., & Arabie, P. (in press). Multidimensional scaling. In M. H. Birnbaum (Ed.), *Handbook of perception and cognition: Vol. 3. Measurement, judgment and decision making.* San Diego, CA: Academic Press.

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

Carver, R. P., & Leibert, R. E. (1995). The effect of reading library books at different levels of difficulty upon gain in reading ability. *Reading Research Quarterly, 30,* 26–48.

Chomsky, N. (1991). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed.), *The Chomskyan turn.* Cambridge, MA: Blackwell.

Choueka, Y., & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities, 19,* 147–157.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics, 16,* 22–29.

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

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Dahl, H. (1979). *Word frequencies of spoken American English.* Essex, CT: Verbatim.

D'Andrade, R. G. (1993). Cultural cognition. In M. I. Posner (Ed.), *Foundations of cognitive science.* Cambridge, MA: MIT Press.

Davies, P. (1971). New views of lexicon. In J. B. Carroll, P. Davies, & B. Richman (Eds.), *Word frequency book* (pp. xli-liv). New York: Houghton Mifflin and American Heritage.

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.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: Johns Hopkins University Press.

Drum, P. A., & Konopak, B. C. (1987). Learning word meaning from written context. In M. C. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 73–87). Hillsdale, NJ: Erlbaum.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), *The Third Text Retrieval Conference (TREC3)*

(NIST Publication No. 500-225, pp. 219–230). Washington, DC: National Institute of Standards and Technology.

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

Durkin, D. (1983). *Teaching them to read.* Boston: Allyn and Bacon.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627–661.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24,* 174–187.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Excel Version 5.0 [Computer software]. (1993). Redmond, CA: Microsoft Corp.

Fillenbaum, S., & Rapoport, A. (1971). *Structures in the subjective lexicon.* New York: Academic Press.

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

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Statistical semantics: Analysis of the potential performance of keyword information systems. *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.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Georgopoulos, A. P. (1996). Motor cortex and cognitive processing. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 507–512). Cambridge, MA: MIT Press.

Golub, G. H., Luk, F. T., & Overton, M. L. (1981). A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. *ACM Transactions on Mathematical Software, 7,* 149–169.

Goodman, N. (1972). *Problems and projects.* Indianapolis, IN: Bobbs-Merrill.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery.* Boston: Kluwer Academic.

Harman, D. (1986). An experimental study of the factors important in document ranking. In F. Rabitti (Ed.), *Association for Computing Machinery 9th Conference on Research and Development in Information Retrieval* (pp. 186–193). New York: Association for Computing Machinery.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554–2558.

Jackendoff, R. S. (1992). *Languages of the mind.* Cambridge, MA: MIT Press.

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

Keil, F. C. (1989). *Concepts, kinds and cognitive development.* Cambridge, MA: MIT Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95,* 163–182.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and reading ability in educational practice and psychological theory. In L. G. Nilsson (Ed.), *Perspectives of memory research* (pp. 325–366). Hillsdale, NJ: Erlbaum.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day English.* Providence, RI: Brown University 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., & Dumais, S. T. (1994). Latent semantic analysis and the measurement of knowledge. In R. M. Kaplan & J. C. Burstein (Eds.), *Educational testing service conference on natural language processing techniques and technology in assessment and education.* Princeton, NJ: Educational Testing Service.

Landauer, T. K., & Dumais, S. T. (1996). How come you know so much? From practical problem to theory. In D. Hermann, C. Hertzog, C. McEvoy, P. Hertel, & M. Johnson (Eds.), *Basic and applied memory: Memory in context* (pp. 105–126). Mahwah, NJ: Erlbaum.

Levy, E., & Nelson, K. (1994). Words in discourse: A dialectical approach to the acquisition of meaning and use. *Journal of Child Language, 21,* 367–389.

Lucy, J., & Shweder, R. (1979). Whorf and his critics: Linguistic and non-linguistic influences on color memory. *American Anthropologist, 81,* 113–128.

Lund, K., & Burgess, C. (in press). Hyperspace analog to language (HAL): A general model of semantic representation. *Language and Cognitive Processes.*

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Ed.), *Proceedings of the 17th annual meeting of the Cognitive Science Society* (pp. 660–665). Pittsburgh, PA: Erlbaum.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua, 92,* 199–227.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Mathematica [Computer software]. (1991). Champaign, IL: Wolfram Research Inc. *Version 5.0*

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

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.

Michaelski, R. (1983). A theory and methodology of inductive learning. *Artificial Intelligence, 20,* 111–161.

Miller, G. A. (1978). Semantic relations among words. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 60–118). Cambridge, MA: MIT Press.

Miller, G. A. (1991). *The science of words.* New York: Scientific American Library.

Moyer, R. S., & Landauer, T. K. (1967). The time required for judgements of numerical inequality. *Nature, 216,* 159–160.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100,* 183–203.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

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

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

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: Erlbaum.

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

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

Osherson, D. N., Weinstein, S., & Stob, M. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* Cambridge, MA: MIT Press.

Pinker, S. (1990). The bootstrapping problem in language acquisition. In B. MacWhinney (Eds.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Erlbaum.

Pinker, S. (1994). *The language instinct: how the mind creates language.* New York: William Morrow and Co.

Pollio, H. R. (1968). Associative structure and verbal behavior. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 37–66). Englewood Cliffs, NJ: Prentice Hall.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Quine (1960). *Word and object.* Cambridge, MA: MIT Press.

Rapoport, A., & Fillenbaum, S. (1972). An experimental study of semantic structure. In A. K. Romney, R. N. Shepard, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (pp. 96–131). New York: Seminar Press.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional nature of sentences. *Journal of Verbal Learning and Verbal Behavior, 17,* 403–417.

Rayner, K., Pacht, J. M., & Duffy, S. A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of Memory and Language, 33,* 527–544.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Loyd (Eds.), *Cognition and categorization* (pp. 28–71). Hillsdale, NJ: Erlbaum.

Schütze, H. (1992a). Context space. In R. Goldman, P. Norvig, E. Charniak, & W. Gale (Eds.). *Working notes of the fall symposium on probability and natural language* (pp. 113–120). Cambridge, MA: American Association for Artificial Intelligence.

Schütze, H. (1992b). Dimensions of meaning. In *Proceedings of Supercomputing '92* (pp. 787–796). New York: Association for Computing Machinery.

Schütze, H. & Pedersen, J. O. (1995). Information retrieval based on word senses. *Fourth Annual Symposium on Document Analysis and Information Retrieval,* 161–175.

Seashore, R. H. (1947). How many words do children know? *The Packet, II,* 3–17.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

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

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89–106). Hillsdale, NJ: Erlbaum.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 546–659.

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

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,* 283–299.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 79–98). Hillsdale, NJ: Erlbaum.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Vygotsky, L. S. (1968). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA: MIT Press. (Original work published 1934)

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

Webster's third new international dictionary of the English language unabridged. (1964). Springfield, MA: G. & C. Merriam Co.

Young, R. K. (1968). Serial learning. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 122–148). Englewood Cliffs, NJ: Prentice Hall.

(Appendix follows on next page)

Appendix

An Introduction to Singular Value Decomposition and an LSA Example

Singular Value Decomposition (SVD)

A well-known proof in matrix algebra asserts that any rectangular matrix (X) is equal to the product of three other matrices (W, S, and C) of a particular form (see Berry, 1992, and Golub et al., 1981, for the basic math and computer algorithms of SVD). The first of these (W) has rows corresponding to the rows of the original, but has m columns corresponding to new, specially derived variables such that there is no correlation between any two columns; that is, each is linearly independent of the others, which means that no one can be constructed as a linear combination of others. Such derived variables are often called principal components, basis vectors, factors, or dimensions. The third matrix (C) has columns corresponding to the original columns, but m rows composed of derived singular vectors. The second matrix (S) is a diagonal matrix; that is, it is a square $m \times m$ matrix with nonzero entries only along one central diagonal. These are derived constants called singular values. Their role is to relate the scale of the factors in the first two matrices to each other. This relation is shown schematically in Figure A1. To keep the connection to the concrete applications of SVD in the main text clear, we have labeled the rows and columns *words* (w) and *contexts* (c). The figure caption defines SVD more formally.

The fundamental proof of SVD shows that there always exists a decomposition of this form such that matrix multiplication of the three derived matrices reproduces the original matrix exactly so long as there are enough factors, where enough is always less than or equal to the smaller of the number of rows or columns of the original matrix. The number actually needed, referred to as the rank of the matrix, depends on (or expresses) the intrinsic dimensionality of the data contained in the cells of the original matrix. Of critical importance for latent semantic analysis (LSA), if one or more factor is omitted (that is, if one or more singular values in the diagonal matrix along with the corresponding singular vectors of the other two matrices are deleted), the reconstruction is a least-squares best approximation to the original given the remaining dimensions. Thus, for example, after constructing an SVD, one can reduce the number of dimensions systematically by, for example, removing those with the smallest effect on the sum-squared error of the approximation simply by deleting those with the smallest singular values.

The actual algorithms used to compute SVDs for large sparse matrices of the sort involved in LSA are rather sophisticated and are not described here. Suffice it to say that cookbook versions of SVD adequate for small (e.g., 100 × 100) matrices are available in several places (e.g., Mathematica, 1991), and a free software version (Berry, 1992) suitable for very large matrices such as the one used here to analyze an encyclopedia can currently be obtained from the WorldWideWeb (http://www.netlib.org/svdpack/index.html). University-affiliated researchers may be able to obtain a research-only license and complete software package for doing LSA by contacting Susan Dumais.[A1] With Berry's software and a high-end Unix work-station with approximately 100 megabytes of RAM, matrices on the order of 50,000 × 50,000 (e.g., 50,000 words and 50,000 contexts) can currently be decomposed into representations in 300 dimensions with about 2–4 hr of computation. The computational complexity is $O(3Dz)$, where z is the number of nonzero elements in the Word (w) × Context (c) matrix and D is the number of dimensions returned. The maximum matrix size one can compute is usually limited by the memory (RAM) requirement, which for the fastest of the methods in the Berry package is $(10 + D + q)N + (4 + q)q$, where $N = w + c$ and $q = \min(N, 600)$, plus space for the $W \times C$ matrix. Thus, whereas the computational difficulty of methods such as this once made modeling and simulation of data equivalent in quantity to human experience unthinkable, it is now quite feasible in many cases.

Note, however, that the simulations of adult psycholinguistic data reported here were still limited to corpora much smaller than the total text to which an educated adult has been exposed.

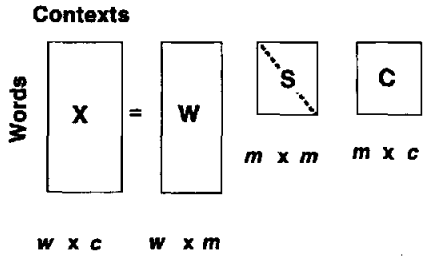

Figure A1. Schematic diagram of the singular value decomposition (SVD) of a rectangular word (w) by context (c) matrix (X). The original matrix is decomposed into three matrices: W and C, which are orthonormal, and S, a diagonal matrix. The m columns of W and the m rows of C′ are linearly independent.

An LSA Example

Here is a small example that gives the flavor of the analysis and demonstrates what the technique can accomplish.[A2] 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. The titles are shown below.

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*

The matrix formed to represent this text is shown in Figure A2. (We discuss the highlighted parts of the tables in due course.) The initial matrix has nine columns, one for each title, and we have given it 12 rows, each corresponding to a content word that occurs in at least two contexts. These are the words in italics. In LSA analyses of text, including some of those reported above, words that appear in only one context are often omitted in doing the SVD. These contribute little to derivation of the space, their vectors can be constructed after the SVD with little loss as a weighted average of words in the sample in which they occurred, and their omission sometimes greatly reduces the computation. See Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) and Dumais (1994) for more on such details. For simplicity of presentation,

[A1] Inquiries about LSA computer programs should be addressed to Susan T. Dumais, Bellcore, 600 South Street, Morristown, New Jersey 07960. Electronic mail may be sent via Internet to std@bellcore.com.

[A2] This example has been used in several previous publications (e.g., Deerwester et al., 1990; Landauer & Dumais, 1996).

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

Figure A2. A miniature dataset of titles described by means of a word-by-context matrix (X) in which cell entries indicate the frequency with which a given word occurs in a given context. (The usual preliminary transformation is omitted here for simplicity.) There are five titles (c1–c5) about human computer interaction and four titles (m1–m4) about mathematical graph theory. Highlighted portions are used to indicate modifications in pattern similarities by dimension reduction between this figure and its dimension-reduced version shown in Figure A4. Here r(human.user) = −.38; r(human.minors) = −.29.

X = W S C'

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

C =

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 A3. The singular value decomposition of the word-by-context matrix (X) of Figure A2, in which cell entries indicate the frequency with which a given word occurs in a given context. Highlighted portions are the values on the first and second dimensions of the component matrices.

$\hat{X}$ =

	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

Figure A4. A least squares best approximation ($\hat{X}$) to the word-by-context matrix in Figure A2 obtained by retaining only the two largest columns and rows from the matrices in Figure A3. Highlighted portions illustrate modifications in pattern similarities by dimension reduction between Figures A2 and A4. In Figure A2 the cell entries indicate the frequency with which a given word occurs in a given context. There are nine titles about human computer interaction (c1–c5) and mathematical graph theory (m1–m4). Figure A3 shows the singular value decomposition (SVD) of the matrix of Figure A2. In this reconstruction, r(human.user) = .94; r(human.minors) = −.83.

the customary preliminary transformation of cell entries is omitted in this example.

The complete SVD of this matrix in nine dimensions is shown in Figure A3. Its cross-multiplication would perfectly (ignoring rounding errors) reconstruct the original.

Next we show a reconstruction based on just two dimensions (Figure A4) that approximates the original matrix. This uses vector elements only from the first two shaded columns of the three matrices shown in Figure A3 (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. 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 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 linear vector arithmetic), my best guess is that word *X* actually appeared 0.66 times in context *Y*."

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. 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 text m4 did contain *graph* and *minors*, the zero entry for *tree* has been replaced with 0.66. By contrast, the value 1.00 for *survey*, which appeared once in text m4, has been replaced by 0.42, reflecting the fact that it is undifferentiating in this context and should be counted as unimportant in characterizing the passage.

Consider now what such changes may do to the imputed relations between words and between multiword 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-dimen-

(*Appendix continues on next page*)

LSA Titles example:

	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

A. Correlations between titles in raw data.

	means	c(1-5)	m(1-4)
	c(1-5)	**0.02**	
	m(1-4)	**-0.30**	**0.44**

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

B. Correlations in first-two principal component space.

	means	c(1-5)	m(1-4)
	c(1-5)	**0.92**	
	m(1-4)	**-0.72**	**1.00**

Figure A5. Intercorrelations (*r*s) among vectors standing for titles in the raw data (A) and the dimension-reduced reconstruction (B). The nine titles are about human computer interaction (c1–c5) and mathematical graph theory (m1–m4). Note how the two conceptually distinct groups have been separated. LSA = latent semantic analysis.

sionally reconstructed matrices (Figures A2 and A4). In the original, *human* never appears in the same context with either *user* or *minors:* they have no co-occurrences, contiguities, or associations as usually construed. The correlation between *human* and *user* is −.38; that between *human* and *minors* is −.29. However, in the reconstructed two-dimensional (2-D) approximation, because of their indirect relations, both have been greatly altered, and in opposite directions: the *human–user* correlation has gone up to .94, the *human–minors* correlation down to −.83.

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 2-D reconstruction. See Figure A5. In the raw co-occurrence data, correlations among the five human–computer interaction titles were generally low, even though all the articles were ostensibly about quite similar topics; half the *r*s were zero, three were negative, two were moderately positive, and the average was only .02. Correlations among the four graph theory articles were mixed, and those between the HCI and graph theory articles averaged only a modest −.30 despite the minimal conceptual overlap of the two topics.

In the 2-D 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 reestimated to be all 1.00, and those between the two contrasting 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. Of course, this is just a tiny selected example. Why and under what circumstances should reducing the dimensionality of representation be beneficial? When, in general, are such inferences better than the original first-order data? We hypothesize that one important case, represented by human word meanings, is when the original data are generated from a source of the same dimensionality and general structure as the reconstruction.

Received December 31, 1995
Revision received July 8, 1996
Accepted August 1, 1996 ■