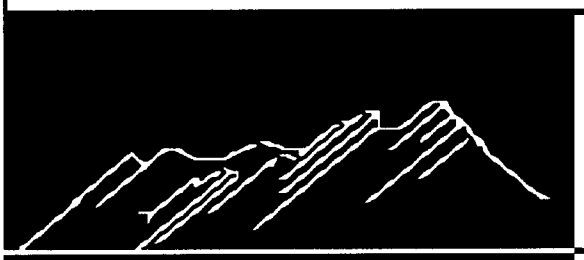
Institute of Cognitive Science



Technical Report

University of Colorado, Boulder

Plato's Problem

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

Thomas K. Landauer

University of Colorado

Susan T. Dumais

Bellcore

Institute of Cognitive Science
University of Colorado
Boulder, Colorado 80309-0344

ICS TECHNICAL REPORT #95-09

Running head: PLATO'S PROBLEM

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

Thomas K. Landauer
University of Colorado

Susan T. Dumais
Bellcore

Abstract

"Plato's problem" is: How do people know as much as they do with as little information as they get? The problem presents itself in many forms, perhaps most dramatically and conveniently for research in learning vocabulary from textual context. A new general theory of acquired similarity, generalization, and knowledge representation, Latent Semantic Analysis (LSA), is presented and used to simulate such learning. By inducing global knowledge indirectly from local co-occurrence data in a representative body of text, LSA approximated the rapid acquisition of meaning similarities by school children. LSA uses no prior linguistic-semantic knowledge or primitive feature similarity relations; it is based solely on a general mathematical learning method that can achieve powerful inductive effects simply by assuming the right (usually high, e.g., 100–350) number of dimensions for its representation of similarity among events. Its possible relations to other theories, phenomena, and problems are sketched.

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

How much do we know at any time? Much more , or so I believe, than we know we know!

—Agatha Christie (1942)

A typical seventh grader knows the meaning of 10 to 15 words today that she didn't know yesterday. She must have acquired almost all of them as a result of reading, because (a) the great 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, thus should have learned less than three new words, not 10 to 15. Apparently, she mastered the meanings of many words that she did not encounter.

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 to complete it.

In this article we suggest a different hypothesis to explain the mystery of excessive learning. The theory 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 simple mechanism of induction, the choice of the correct dimensionality in which to represent similarity between 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.

Introduction

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 approached the natural rate for school children, 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 the empirical result proves 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 above and elaborate below, 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 at least 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 matching—that greatly amplifies its learning ability, allowing the model to correctly infer indirect similarity relations only implicit in the temporal correlations of experience. The model exhibits human-like 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, 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 phenomena to which it may be germane. The mechanism lends itself, among other things, to a reformulation of associational learning theory that appears to offer explanations and modeling directions for a variety of cognitive phenomena. It might also be construed as an organizational mechanism for implicit memory. One set of phenomena that we discuss 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 follow a slightly unorthodox manner of exposition. While we 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 try 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 truth 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," "the poverty of the stimulus," or, in another guise, "the problem of the expert," 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, Quine), psychologists (e.g., Hunt, Osherson, Rumelhart & McClelland; Shepard, Vygotsky), linguists (e.g., Chomsky, Jackendoff, Pinker), computation scientists (e.g., Feldman, Gold, Hinton, and Sejnowski), and combinations thereof (Holland, Holyoak, Nisbett, & Thagard, 1989) have wrestled with the problem in many guises. Quine (1960), following a tortured history of philosophical analysis of scientific truth, calls 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 recognized 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. School children learn about 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 learn 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 strong and specific innate foundation, a set of general rules and predilections which need parameter-setting and filling in, but not acquisition as such, from experience. While 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, and how it works. Unfortunately, the rest of the instinctivist answers are 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 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 posits that the nervous system has evolved general functional relations between monotone transductions of input values and the similarity of central interpretive processes. On average, he maintains, 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 law for stimulus generalization is empirically correct, or nearly so, for a considerable range of low-dimensional, psychophysical continua, and for certain functions computed on behaviorally measured relations such as choices between stimuli or judgments of 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 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, 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, 1986). 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 appoach 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 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 (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, like 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.

Something of the sort is the apparent aim of Chomsky's program for understanding the acquisition of grammar. He supposes that the mind contains a prototypical framework, a set of kinds of rules, on which any natural language grammar can be built, and that being required to obey some one of the allowable sets of rules sufficiently constrains the problem that a child can solve it; a small amount of evidence will suffice to choose between the biologically possible alternative grammars. Of what the presumed primordial, universal, abstract grammar consists remains unsettled, although some of its gross features have been described. How experiential evidence is brought to bear in setting its options also has yet to be well specified, although developmental psycholinguists have provided a great deal of relevant evidence. Finally, the rules so far hypothesized for "universal grammar" are stated in sophisticated mentalistic terms, like "head noun," that beg for reduction to a level at which some logical or neural computation acting on observables or inferables can be imagined for their mechanism.

A similar tack has been taken in attempting to explain the astonishing rate of vocabulary learning—some 7 to 10 words per day—in children during the early years of preliterate language growth. Here, theorists such as E. Clark (1987), Carey (1985), 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 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 soluble on the data available. So, for example, if biology specifies a function on wavelength of light 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. A syntax learner who can assume that verbs either always precede nouns, or always follow them, need only learn which; a word-referent learner who can assume that no two words refer to the same object, when presented with an as-yet unnamed object and an as-yet unknown word can guess with reasonable safety that they are related to each other.

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 usage; ask 100 people what to title a recipe or name a computer command and you get almost 30 different answers on average—see Furnas, Landauer, Dumais, & Gomez, 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 specific and relevant only to human language, making their postulation consistent with a strong instinctive and modular view of mental processes. In Pinker's (1994) recent pursuit of this reasoning he is led to postulating, albeit apparently with tongue somewhat in cheek, no less than 15 different domains of human knowledge, each with its own set of specific innate-knowledge constraints. Is it likely that such a panoply of domain-specific innate knowledge could have arisen over less than a million years of Homo Sapiens evolution? Or is some more general set of constraints, in spirit more like those proposed by Shepard, at work throughout cognition? One potential advantage of more general cognitive constraints is that they might make possible derived sets of higher-order constraints based on experience, which could then underwrite induction in relatively labile domains of knowledge such as those aspects of culture invented slowly by earlier generations but learned quickly by later ones.

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. The rest of the problem involves three general issues. The first is 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?

The second issue is methodological; how to get an answer to the first question, how to determine whether a specified combination of constraints when applied to natural environmental input would solve the problem, or perhaps better, determine how much of the problem it would solve. We believe that the best available strategy for doing this 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 many facts, each weakly related to 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 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.

The third problem is to determine whether a postulated model corresponds to what people actually do, whether it is psychologically valid, and whether the constraints it uses are the same ones on which human achievement relies. As we said earlier, showing that a particular constraint (e.g., avoidance of synonyms) exists in a knowledge domain and is used by learners is not enough unless we can show that the constraint sufficiently helps to solve the overall inductive problem over a representative mass of input. 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 when embedded in a simulation model running over a natural body of data the same constraints approximate natural human learning and performance. 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, while 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 we describe are plausible, to reduce its otherwise magical appearing 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 cyclical fashion. First, we explain the underlying inductive mechanism of dimension matching on 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 then, finally, return to a more detailed and rigorous presentation of the model and its applications.

Suppose that two people who can only communicate by telephone are trying to pass information. Person A, sitting high on a ridge and looking down at the terrain below estimates the distances separating three houses: one green, one blue, and one yellow house. She says that the blue house is 5 units from both the red and yellow houses, and the red and yellow houses are separated by 8 units. Person B uses these estimates to plot the position of the three houses, as shown in the top portion of Figure 1. But then, Person A says, "Oh, by the way, they are all on the same straight, flat road." Now Person B knows that Person A's estimates must have contained errors and revises his own estimates in a way that uses all three distances to improve each one (to 4, .5, 4.5, and 9) as shown in the bottom portion of Figure 1.

Insert Figure 1 about here

Three distances among three objects are always consistent in 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 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 method

works the other way around as well. If the distances had been generated from a two- or three-dimensional array—for example, the road was curved or curved and hilly—projecting the estimates onto a straight line would have distorted their original relations and added error rather than reducing it.

Sometimes researchers have considered dimension reduction methods to be techniques for merely reducing computational complexity or for smoothing, that is for reducing random error or approximating missing values by averaging and interpolating (e.g., Church & Hanks, 1990; Grefenstette, 1993; Shutze, 1992). Dimension reduction does have those advantageous properties, of course, but dimension matching—choosing the right dimension, when appropriate—can be a much more powerful step. Representing data in a dimensionality different from its source, either too few or too many, can introduce not just noise but systematic errors.

Let us now construe the semantic similarity between two words as a distance: the closer the distance the greater the similarity. Suppose we also assume that the likelihood of two words appearing in the same window of discourse—a phrase, a sentence, a paragraph, or what have you—is inversely proportional to their semantic distance, that is directly proportional to their semantic similarity.[1] We can then estimate 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. Worse yet, estimates for most pairwise relations would be completely missing, not only because of thin sampling, but also because real language may use only one of several words of near synonymous meaning in the same passage (just as only one view of the same object may be present in a given scene). If the internal representation of semantic similarity is constructed in a limitless number of dimensions, there would be nothing more we could do with the data. However, if the source of the discourse was a mind in which semantic similarities were represented in $\underline{k}$ dimensional space, then we might be able to improve our initial estimates of pairwise similarities and accurately estimate the

similarities among pairs never observed together, by fitting them as best we could into a space of the same dimensionality. This method is closely related to familiar uses of factor analysis and multidimensional scaling, and to unfolding (Coombs, 1964), but using a particular kind of data and writ 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 will result from dimension matching 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 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 in some region of the space. Locally, over short time spans, similarities among output words would reflect, at least weakly, their distances 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 context (e.g., a paragraph). However, because there is a large number of words in any natural language, and a relatively small amount of received discourse, such information would surely be inadequate. For example, it is quite likely that two words with frequencies of one in a million will have never been experienced near each other even though they have related meanings. However, if the receiving device sets out to represent the results of its statistical knowledge as points in a space of the same dimensionality as that from which it was

generated, it is bound to do better. How much better will depend, as we've already said, on matters that can only be settled by observation.

Except for some technical matters, such as the similarity metric employed, 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, the model assumes that the psychological similarity between any two words is reflected systematically in the way they co-occur in small subsamples of language. The model assumes that the source of language samples produces words in a way that ensures a relation between semantic similarity and output distance that will allow recovery of the semantic similarities by fitting all observed pairwise similarities into a common space of high, but not unlimited, dimensionality.

As in the house mapping and geometric examples, the assumed number of dimensions cannot be too great or too small for such a trick to work. That is, to utilize the extra information inherent in the dimensional constraint, the receiver must know or discover the dimensionality of the source. Not knowing this dimensionality a priori, we varied the dimensionality of the simulation model in our studies to determine what produces the best results.[2]

More 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 psycholinguistics 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 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" mental

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 provide a more complete description of LSA as a mathematical model below 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 that represent unitary event types by columns that, in turn, represent 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, closely akin to factor analysis, that allows event types and individual contexts to be rerepresented as points or vectors in a high-dimensional abstract space. The final output is a representation from which we 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 local associations between a stimulus and other temporally contiguous stimuli, or, equivalently, as associations between stimuli and the contexts or episodes in which they occur. The stimuli or event types may be thought of as unitary chunks of perception or memory. (We describe a hypothetical unitization process later that is, in essence, a hierarchical recursion of the LSA representation.)

The first-order process by which initial pairwise associations are entered and transformed in LSA resembles classical conditioning. 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 that 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 association, for example, between a word and closely following words. Indeed, Lund and Burgess (1995; in press) explored a related model 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 step 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 (e.g., the association of A and B will be greater if both appear in two different contexts than if they each appear twice in the same context). In a second transformation each of these cell entries is divided by the entropy for the event type, $-\sum p \log p$ over all its contexts. Roughly speaking, this step accomplishes much the same thing as conditioning rules like those described by Rescorla and Wagner (1972), in that the step discounts the effect of a pairing by the frequency of occurrence of the same events unpaired, thus making the association better represent the informative relation between of event types rather than the mere fact that they occurred together.

It is interesting to note that automatic information retrieval methods (including LSA when used for the purpose) are improved greatly by transformations of this general form. 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 they draw the words in 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 clear and precise way. The initial matrix after transformation to log frequency/entropy represents the product of the local or pairwise processes.[3] The subsequent analysis and dimension 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 a conditioning or associative processes. However, the model's next steps, the singular value decomposition and dimension reduction, are not contained in any extant theory of learning, although something of the kind may be hinted at in some modern discussions of conditioning, and is latent in many neural net and spreading activation architectures. What this step does is convert 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 rerepresentation 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 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 simple chaining rules have not been notably successful. If associations correspond to distances in space, as supposed by LSA, simple 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 0 and 4. 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 stable compromise configuration (physical objects, weather systems, and Hopfield nets do this too). It is easy to see that the resulting relation between any two representations depend not only on direct experience with them but also with everything else ever experienced. No single representation will be an island. Although the current mathematical implementation of LSA doesn't 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, will produce a faithful reflection of human knowledge.

Finally, to anticipate what will be developed below, 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 LSA optimizes the prediction of the presence of all other events

from those currently identified in a given temporal context, and does so using all relevant information it has experienced.

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

A Neural Net Analog of LSA

We describe SVD more fully, but still informally, shortly (and in somewhat greater detail in the Appendix). For those familiar with neural net models, we offer first a rough equivalent in that terminology. Conceptually, the LSA model can be viewed as a simple but rather large three-layer neural net. It has a layer-one node for every word (element) and a layer-three node for every text window (episode) ever encountered, several hundred layer–two nodes, and complete connectivity between layers one and two and between layers two and three. (Obviously, one could substitute other identifications of the elements and episodes.) The network is symmetrical; it can be run in either direction One chooses some large number of middle-layer nodes, then maximizes the accuracy (in a least squares sense) with which activating any layer-three node activates the layer-one nodes that are its elementary contents, or vice versa. The conceptual representation of either kind of event, a unitary episode or a combination of episodic component elements, is a pattern of activation across layer-two nodes. All activations and summations are linear.

The Singular Value Decomposition (SVD) Realization of LSA

The principle virtues of SVD for this work 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 is only one case of a class of potential models that one would eventually wish to consider, a case which uses a 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 (e.g., certain aspects of grammar and syntax that involve ordered and hierarchical relations rather than unsigned distances). However, what is, most interesting at this point is how much it does in its present form.

Singular Value Decomposition (SVD)

A brief overview the mathematics of SVD is given in the Appendix. For those who wish to skip it, we note that SVD is the general method for linear decomposition of a matrix into independent factors, 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 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; 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 special case of Eigenvalue—Eigenvector analysis, closely related to principal components decomposition and, in a more general sense, to multidimensional scaling.)

To implement the model concretely and simulate human word learning, SVD was used to analyze 4.6 million words of text taken from Grolier's Academic American Encyclopedia (19xx), a work intended for young students. This encyclopedia has 30,473 articles. From each article we took a sample consisting of the whole text or its first 2,000 characters, whichever was less, for a mean text sample length of 152 words, roughly the size of a rather long paragraph. The text data were cast into a matrix of 30,473 columns (each column represented one text sample) by 60,768 rows (each row represented a unique word type that appeared in at least two samples). The cells in

the matrix contained the frequency with which a particular word appeared in a particular text sample. The raw cell entries were first transformed to {ln (cell frequency)/ entropy of the word over total collection}. 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 generated a vector of 300 real values to represent each word (see Figure 2). The similarity of words was usually measured by the cosine between their vectors.[4]

Insert Figure 2 about here

We postulate that the power of the model comes from dimension reduction. Here's 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 elaboration in yet another way, a change in the value of any cell in the original matrix can, and usually will, 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.

Neurocognitive Plausibility

We, of course, intend no claim that the mind or brain actually computes a singular value decomposition 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 (or 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 fft the way a computer does, only that the nervous system is sensitive to and produces a result that reflects the spectral composition of the input. For LSA, hypotheses concerning how the brain's parallel neural processing might produce an SVD-like result remain to be specified, although it may not be totally vacuous to point out that the brain's interneuronal communication processes are effectively vector multiplication processes between axons, dendrites, and cell bodies, and that the neural net model's popularly used to simulate brain processes can be recast as, and indeed are often calculated as, matrix operations.

Testing the Model

Four pertinent questions were addressed by simulation: Could such a simple linear model acquire knowledge of human-like word meaning similarities to a significant extent if given a large amount of natural text. Supposing it did, would its success depend strongly on the dimensionality of its representation? And how would its rate of acquisition compare with that of a human reading the same amount of text? Finally, how much of the model's knowledge would come from indirect inferences that combine information across samples rather than directly from the local contiguity information present in the input data?

In answer to the first question we begin with results from the most successful runs, which used 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 old items from the synonym

portion of the Test of English as a Foreign Language (TOEFL), kindly provided by Educational Testing Service. 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 and/or the correct alternative, for which it was given a score of .25). The model got 51.5 correct, or 64.4% (52.5% corrected for guessing). By comparison, a large sample of students from non-English speaking countries who took tests containing these items averaged 51.6 items correct, or 64.5% (52.7% corrected for guessing). LSA's pattern of cosines over the incorrect alternatives of the items correlated, on average, at .44, with the relative frequency of student choices, close to the expected value for the correlation of a randomly selected student with whole group.

Thus, the model quite 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 the model supposes that the correct choice of dimensionality is important to success. To determine whether it was, the simulation was repeated using a range of dimension

numbers. Two or three dimensions, as used, for example in many multidimensional scaling attacks on word meaning (e.g., Fillenbaum & Rapoport, 1971; Rapport & Fillenbaum, 1972) and in the Osgood semantic differential (1957) were totally insufficient, resulting in only 13.3% correct answers when corrected for guessing. More importantly, using too many factors also resulted in poor performance. With no dimension reduction at all, that is, using cosines between rows of the original 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. Thus, choosing the dimensionality of the representation well approximately tripled the number of words the model learned (see Figure 3; note that the x-axis is not a linear scale but a set of selected values).

Insert Figure 3 about here

Comparing the Learning Rate of LSA to Humans and Assessing Its Reliance on Induction

Next, 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 do well on 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 humans possess. Unfortunately, we can't use the ETS normative data directly for this comparison because we don't know how much English their test takers had read, and because, unlike LSA, the ETS test takers were primarily second language learners.

For similar reasons, while we have shown that LSA makes use of dimension 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 School Children's Vocabulary Acquisition

LSA gains its knowledge of words by exposure to text, a process at least partially analogous to reading. How much vocabulary knowledge do humans learn from reading and at what rate? We expand here somewhat on the brief summary given in the prologue. 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 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 (see 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. 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 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 gotten right by knowing some other morphologically related word, and adjusted the results accordingly. The resulting estimate was around 100,000 words. This is at the high end of published estimates. A lower, 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 of which are common (Walker & Amsler, 1986).

By simple division, knowing 40,000 to 100,000 words by 20 years of age means adding an average of 7 to 15 new words a day from age two onward. 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 to 15 per day), with some years in late elementary school showing more rapid gains than the average (Nagy & Herman, 1987; Smith, 1941). In summary, it seems safe to assume that, by the usual measures, normal fifth to eighth grade students acquire the meaning of somewhere between 10 and 15 new words per day.

As mentioned earlier, the acquisition of almost all these new word meanings must depend on reading. The proof is straightforward. The number of different word types in spoken vocabulary is much smaller than that in written vocabulary; the words that individuals hear in daily intercourse with family and friends probably account for less than one quarter of an adult's comprehension vocabulary. Most school children spend more than a third of their waking hours in front of television sets, and the vocabulary of television discourse is even more limited. Moreover, because the total quantity of heard speech is large, and spoken language provides superior cues for meaning acquisition, such as perceptual correlates, pragmatic context, gestures, and outright feedback and disambiguation 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 toward the tens of thousands used in speech by an average adult (Seashore, 1947). Finally, little vocabulary is learned from direct instruction. Most schools devote 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 minutes per day reading in school and another 15 out of school reading books, magazines, mail, and comic books (Anderson, Wilson, & Fielding, 1988; Taylor, Frye, & Maruyama, 1990). If we assume 30 minutes per day total for 150 school days and 15 minutes per day for the rest of the year, we get an average of 21 minutes per day. At an average reading speed of 165 words per minute (Carver, 1990), which may be an overestimate of natural, casual rates, and a nominal paragraph length of 70 words, children read about 2.5 paragraphs per minute, about 50 per day. Thus, while reading, school children are adding about one new word to their recognition vocabulary every 2 minutes, 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 and their definitions each day and expect them to permanently retain 10 to 15 associations after a single 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 (Eley, 1989; Nagy et al., 1985). About one word in 20 paragraphs makes the jump, a rate of 0.2 words per paragraph read. At 50 paragraphs read per day, children would acquire only 2.5 words per day.

Thus, experimental attempts intended to produce accelerated vocabulary acquisition have attained less than one half the natural rate, and measurements made under realistic conditions find at best one-fourth the normal rate. These results lead 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 now 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 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. Because these effects depend on both 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-aged 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 (Carroll, Davies, & Richman, 1971) was based. By assuming that learners would acquire knowledge about the words in the 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.

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 on our encyclopediac 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 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, respectively. 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, 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.[6] 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 ratio measure, computed by dividing the cosine between the stem word and the correct alternative by the standard deviation of cosines between the stem and the three incorrect alternatives. That is, for each test item separately, we found the cosine between the stem and each alternative, calculated the standard deviation for the three incorrect alternatives, and divided it into the cosine for the correct alternative. 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 (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, $p = .001$). Because of the interaction, the absolute sizes of the two overall effects taken separately (i.e., averaged over the

other variable) are not interpretable except to demonstrate the existence of each. These effects are illustrated in Figure 4, along with logarithmic trend lines for T within each level of S.

Insert Figure 4 about here

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 will contribute, one must know how many other text samples with and without that word the learner or model has previously met.

In the last analysis step, we asked, for an average 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 school child's vocabulary learning history. We then estimated 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 .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 .15 words. Multiplying by a nominal 50 samples read, we get a total vocabulary increase of 10 words per day. Details of this analysis are given next.

Details of Simulation of Total Vocabulary Gain

For this purpose we could have rerun the analysis again for a chosen experience level, say an amount of text corresponding to what a seventh grader would have read, and for the frequency of each word that such a child would have encountered. However, such an approach would have two disadvantages. One would have been simply its excessive computation time. More important, in

principle, is that such a procedure would introduce undesirable sampling variability (notice, for example, the somewhat unsystematic variations in slope across the random samples constituting S levels in Figure 4). Instead, 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. For the purpose at hand, this model need only be correct at a descriptive level, providing a single formula based on a collection of data across representative points and predicting effects at all points in the joint effects space. The formula below does a good job:

$$z = a * (\log b\ T) * (\log c\ S) \qquad \text{(Equation 1)}$$

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 iterative solver program). Its predictions are correlated with observed z with $\underline{r} = .98$. 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 at $\underline{r} = .89$ with observed percent correct.

We were then ready to make the desired estimates for simulated children's learning. To do so, we needed to determine for every word in the language (a) the probability that a word of its frequency appears in the next text sample a typical seventh grader will encounter, and (b) the number of times she would have encountered that word previously. We then estimated, 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 gain from reading a single text sample.

The Carroll et al. data give the frequency of occurrence of each word type in a representative corpus of text read by school children. 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. The two samples might still have differences in the shape of their distributions (e.g., in the ratio of rare to common words or in the way related words aggregate over paragraphs because of content differences). However, because the effects of such differences on the model's parameter estimates would be small we ignored them.

To simulate the rate of learning for a late grade school child, we assumed that she would have read a total of about 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 Carroll et al. into 37 frequency bands (0,1,2...20, 21–25, and roughly logarithmic thereafter to > 37,000) and for each band set S equal to an interpolated central frequency of words in the band.[7] 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.[8]

The result was that the estimated direct effect was .0007 words gained per word encountered, and the indirect effect was a total vocabulary gain of .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 .0007 * 70 (approximately the number of words in an average paragraph) + .15 = .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.

Before 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 school children. 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 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 152 words, and we have equated them with assumed 70-word paragraphs read by school children because our belief 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 and paragraphs of text read by children. Informal results from other work, mostly in information retrieval, but also in pilot studies for this simulation using smaller mixed text sources, have found roughly the same results over sample sizes ranging from 20 to a few hundred words, generally following a gently nonmonotonic inverted-U function, and it is unlikely that the amount of learning would increase significantly, if at all, above 70 words per sample. Nevertheless, this is an issue that cannot be fully resolved until the simulations can be run starting with SVD on the same text as children read. For the worst case, recomputing the estimates of effects per total words rather than per sample or paragraph; that is, assuming that amount of learning about a particular word in a sample is linear with the total number of words in the sample,

yields an estimate of six new vocabulary words per day for the simulation, still much greater than laboratory estimates of context-based learning.

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 cut-off was used because the available machine-readable text had article separators but no consistent paragraph or sentence indicators. This format resulted in the inclusion of a large number of short samples, 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 size. Several alternatives suggest themselves. We speculate below that a variety of different window sizes might be used simultaneously. 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 and sentences and other coherent segments rather than arbitrary isolated pieces. Thus, although LSA learning does not appear to be sensitive to small differences in the context window size, this factor, too, was not optimized in the reported simulations as well as it may be in human reading.

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. While 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 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 carries. Experts on reading instruction (e.g., Durkin, 1979; Drum & Konopak, 1987) 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) 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 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 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, or 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 dimension-matching mechanisms.

Conclusions From the Vocabulary Simulation Studies

There are four important conclusions to be drawn from the results we 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. At least half of LSA's knowledge, and probably three-quarters, 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 text and has a moderately better system for input parsing, the human induction system can do the otherwise apparently mysterious learning that it does of the same relations without recourse to language-specific innate knowledge.
4. Because of its inductive properties, the rate at which LSA acquires this knowledge from text is much greater than the rate at which it gains knowledge of the particular words present in text to which it is exposed, just as is the case for school children when reading.

Let us return for a moment to the apparent paradox of school children 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 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 are gotten right afterwards. This method might be less of a problem if all words in the text being read were tested and a large enough sample were measured. But what usually has been done instead is to select for testing only words likely to be unknown before reading. This simplifies testing, but introduces a potential bias. In contrast, the LSA simulations computed an increment in probability correct for every word in the text (as well as every other 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 known perfectly. 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, could mimic other aspects of lexical knowledge as well.

Generalizing the Domain of LSA

There is no reason to suppose that the mind uses dimension matching only to induce the similarities involving words. Many other aspects of cognition would also profit from a means to extract more knowledge from local association data. While the full range and details of LSA's implications and applicability await more research, we provide examples of promising directions, phenomena for which LSA 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. While it is probably impossible 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 solely 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. 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 need ever have been directly associated 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 relations other than gross similarity. 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. While 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 pre-established to be something like dog, animal, object, furry, cute, fast, ears, and so forth, it is much less mysterious that a few contiguous pairings of the word rabbit with scenes including the thing itself will 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 (but see conjectures below on unitization), and recognition parts of the problem, but even here LSA may eventually help by providing a powerful way to generate and represent learned and indirect similarity relations among perceptual features. Nevertheless, 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 teenager that the Web is awesome, 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 (whatever that means) on breadfruit and poi (whatever those are). More-or-less correct usage often precedes referential knowledge, which itself can remain vague but connotatively useful. Thus, the frequent arguments over the meaning of words and the livelihood of lexicographers and language columnists who educate us about words we already partially know. 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 gavaggi problem. He asks us to imagine a child who sees a scene in which an animal runs by. An adult says "gavaggi." What is the child to think gavaggi means: ears, white, running, 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 will acquire similarity to each other but not to the rest. (This is illustrated in the Appendix figures 3 and 4, which we urge the reader to examine.) This inductive process solves the part of the problem 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 gavaggi? Well, of course, these might actually be what's meant. But by LSA's inductive process, component features of legs, ears, fur, and so forth, will either before or later all be 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 importantly, by occasions in which they do not occur at all. Thus, the new

object in view will not be 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 will be changed and made yet more similar to each other and to any word selectively used in their presence. Moreover, by the hypothetical higher order process alluded to earlier, the whole gavaggi, on repeated appearance, may take on unitary properties even though it looks somewhat different each time.

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 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 much the same 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 an 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 will her novel name for the child's similar private experiences.

Let us turn now to a description of the hypothetical process by which unitary event-type nodes might be generated, after which we come back to some dependent issues in semantics on which we present some data.

Association, Perceptual Learning, and Chunking

In this section we take the notion of the model as a homologue of associative learning several 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 such as those to be suggested next demands more data, computational resources, and time than has been available. Nevertheless, a sketch of possible implications and extensions show how the dimension-matching inductive process might help to explain a variety of important phenomena that appear more puzzling without it, and suggest new lines of theory and investigation.

After the dimension 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 dimension-matching mechanism to produce similarities and generalizations among psychological entities of many kinds, such as stimuli, responses, percepts, concepts, memories, ideas, images, and thoughts. A few examples follow.

Because the representation of all kinds of entities is the same and association is mutual, the overall growth of knowledge produces a complex structure by a recursive process in which new units are built out of old ones. One way to imagine the process follows. Suppose the naive mind (brain) constantly generates new temporal context vectors to record passing episodes of

experience. We may think of such vectors as akin to new nodes in a semantic or neural network, in that they represent their input and output as weights on a set of elements or connections. We assume that in the real-time dynamics of the system nodes are equivalent to LSA vectors that are activated, in the sense that they are temporarily capable of having their connection weights (their vector element values) altered, and of activating other nodes in proportion to the similarity of their vectors. Further, we assume that the temporal durations of the activity of these nodes, and thus of the episodes they come to code and represent, are distributed over a range, either because of inherent life-span differences or as a result of the dynamics of their interaction with other nodes.

The mind also receives input-vector activations from primitive perceptual processes. Every primitive perceptual vector pattern will, perforce, become locally associated with one or more temporal context nodes. Because of the dimension-reduced representation, context node vectors will acquire induced similarity. This, in turn, will mean that particular context nodes will be reinitiated by new, now similar, primitive percepts (e.g., oriented visual edges and corners that, by induction, belong to the same higher order node, that is, to a representative of events of longer duration, and by new, now similar, higher order vectors). These higher order vectors will themselves form local associations with other higher order vectors representing contexts of both longer and shorter durations, and so forth.

So far, this process may seem little more than the workings of a complex associative network. What makes it different is the glue of dimension-matching induction—that every node is related to every other through common condensed vector representations, not just through independently acquired pairwise node connections and their composite paths. This process gives perceptual and observational learning, and the spontaneous generation of abstractions such as chunks, concepts and categories much greater force and flexibility.

Originally meaningless node vectors would take on increasing repeatability, originate from a variety of sources, and come to represent concepts of greater and greater abstraction—concepts that

stand at once for the elementary vectors whose joint occurrence composed them, other elementary vectors with induced similarity, context vectors to which they have themselves been locally associated, and context vectors with induced similarity to those. But because each node will tend to reactivate ones similar to it, and node vectors of longer durations will come to represent more related components, local hierarchies and partial orders will be statistically common. One aspect of this hypothetical process is a mechanism for the creation of unitary "chunks," vectors representing associations and meanings of arbitrarily large scope and content, the unitization process to which we referred above.

This notion of a hierarchical associative construction process in which larger concepts are built of smaller concepts, which are built of smaller concepts, and so on is not especially novel. However, the proposed mechanism by which lower order elements combine into higher order representations is. The new mechanism is the condensation of all kinds of local correlational evidence into a common representation containing vectors of the same kind at every conceptual level. One result of this process is that all elements at all levels have some degree of implicit association or similarity with every other. The degree of similarity will tend to be greatest with other elements of the same or similar life spans. Thus, an elementary speech sound will have close similarity to frequently following speech sounds and to ones that could occur in its contextual place and to syllables of which it is or could be a part, but will also be similar to varyingly lesser degrees to every episode of its owner's life. Almost any fact, say an old friend's name or an autobiographical event, might be brought to mind by an almost unlimited number of things related to it in any way at all—for example, by Proust's tea-soaked madeleine—and multiple indirectly related and individually weak associates would combine to yield stronger recollections.

Because of the mathematical manner in which the model creates representations, a condensed vector representing a context is the same as the 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 elements also has a vector representation in the space, which in turn gives rise to similarity and generalization effects among new event complexes in an identical fashion to those for two old elements or two old contexts. In some examples we give later, the consequences of representing larger segments of experience as the vector sum of the smaller components of which they are built will be illustrated. For example, we show how the condensed (mathematical centroid) representation of a sentence or a paragraph predicts comprehension of a following paragraph, whereas its sharing of explicit words does not. We provide examples in which the condensed representation for a whole paragraph determines which of two words it is most similar to, while any one word in it does not.

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 from experience, while the context representations are "episodic," unique combinations of events that occurred only once. Yet both are represented in the same space, 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. And the retained knowledge of the context paragraph in the form of a single average vector is itself a representation of "gist" rather than surface detail.

One unsolved problem in this conceptualization is how discrete, semantically unitary nodes like words are created. In the LSA simulations, each word is assigned its own row, and each occurrence of the same word (i.e., letter string) is duly tallied against that row. On the other hand, each context is treated as unique; no two are ever assigned the same column. Of course, the strict separation into separate types of frequently repeating and never repeating events, imposed ex cathedra in our simulations, is not a likely property of nature. It is also not a mathematical necessity in LSA. However, for the correlation-based condensation to be applied, there must be some sense

in which there is repetition of units so that frequency of local co-occurrence can be exploited. Thus, there must exist some way in which highly similar experiences attain unitary status by which a particular representational vector can be part of and be modified by different occasions.

The node notion introduced above, and the idea that nodes are activated by the simultaneous activity of other nodes corresponding to similar vectors, might provide the underpinning for such a mechanism. But we still need to understand how conglomerates of experience of potentially unlimited variability turn into discrete unitary wholes. Obviously, this has much to do with the question of symbolism, a matter much worried over the last few years by neural-net theorists and linguistics. So far, we have suggested that originally meaningless nodes gain first the ability to be rekindled by the events they witnessed and by things similar thereto, and later, inductively, by more and more of the same. Although a mechanism remains to be determined, it seems plausible that the effect is realized through positive feedback in which a node once partially defined—its vector initially set—will find itself more often reactivated by and in turn reactivating related nodes in naturally connected event streams and scenarios, iteratively concentrating its defining vector and separating it from others. Such a process would, conceivably, generate words at one level of granularity, situational schemas at another, and still other representational units. However, it is not clear whether feedback alone would suffice without some sort of nonlinear competitive mechanism.

In our initial simulations we used just two extremes of the repeatability continuum. Words, possibly more than any other entity (but syllables, letters, and mothers' faces may be other candidates) have the opportunity to become unified. In language, the learner is capable of producing the self-same perceptual events that it needs to categorize. It can hear and say variants of sounds until it has agreed with itself and its neighbors by the communicative consensus process outlined in the beginning of this article. Conceivably, this ability was the touchstone of Homo's invention of language. The story goes like this: An LSA-like inductive mechanism originally

evolved to abstract and represent external physical objects and events that are only partially and poorly repeatable. Applying it to easily varied and repeated motor outputs that produce easily discriminable perceptual inputs, such as hand, arm, and vocal gestures, would almost automatically result in intra- and interindividual agreement on the usage, referential or expressive meaning, and similarity of the gestures. The obvious adaptive advantage of agreed gestures would then have guided expansion of the basic mental capacity and of better organs of expression. This logic and evolution would, perhaps, have applied first to mimetic gesture using hands freed by upright posture and trained by tool use, and later to primitive speech sounds, oral words, letters, and written words. The fact that new sign languages, creoles, and written symbol systems have developed spontaneously in many relatively isolated communities, sometimes in only a few generations (the American Plains Indians offer good examples), is a consistent bit of evidence that agreement on the usage of communicative elements can be rapidly accomplished, as LSA would support.

Unfortunately, however, we are not quite done with the node unification issue. Another aspect of it concerns the nature of the node-vectors that are formed, in particular whether a single vector represents every discriminated event type, a range of variations on an event type, or several different event types. For example, is there just one vector corresponding to the word bank, one for bank the institution plus the act of depositing money therein, a second for banks as riversides, and still another for shoveling operations? Or is there a separate vector for every discriminated meaning of bank? On the other hand, can the same event type be represented by more than one vector, or does some competitive process assure that every vector carries a different significance? In nature many different words, if used in the same contexts, will be understood to refer to the same object or event (synonomy), although whether a given individual would use a different word without intending a different meaning can be questioned. And almost every word is understood in different contexts to refer either to more than one quite different thing or carries a somewhat

different meaning (polysemy). Possibly, the number and relative dominance of different meanings for a symbol node and different symbols nodes for similar events arises simply from a combination of sampling frequencies and the inherent passive competitive effects of linear condensation; the fact that common objects tend to have more applicable words and common words evoke more variable contextual meanings suggests something of the sort. However, the issue is very much open.

The dual problem of synonomy and polysemy confronts the LSA model realized in the present simulations in an interesting way. To repeat, by fiat, row entities (words) are represented as repeating units; every time the same spelling pattern is encountered it is assigned the same node (row). Thus, a word is not allowed to correspond to more than one meaning vector. If a spelling pattern has occurred in several dissimilar contexts, LSA is forced to choose for it one vector that represents their weighted average rather than two or more that approximate different senses appropriate to the different contexts. However, as noted earlier, if we were to assign each separate occurrence of a word—each of its tokens—a new row, there would be no first-order associational data on which to induce the condensed representation. There are two solutions. The one we currently favor (suggested to us by Walter Kintsch) is that separate senses are <u>not</u> separately represented in memory but are generated dynamically as evanescent points in meaning space. The lexicographer's differentiation and description of them is, then, just a convenient classification of the common contexts in which a given word appears and the way in which its presence therein changes the meaning of the whole. The second possibility is that there are intermediate levels of representation, additional nodes identical to neither words nor individual contexts. The number of such nodes would have to be limited, else the effective constraints of dimension matching would be lost. Moreover, a new, perhaps dynamically competitive, notion of the connection of a word to its contexts would have to be introduced. All this goes well beyond the present discussion and will not be taken further, but the issues and ideas involved will be revisited later when we consider contextual disambiguation.

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 hours 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. While these extrapolations should not be taken too 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. This amplified learning is a part of long-term memory as ordinarily conceived, although indirect effects would also be expected in the capacity of working memory, as recently suggested by Ericsson and Kintsch (1995).

The growing value of unconscious induction is a familiar introspective experience. A psychology professor can automatically extract and extend the knowledge contained in a psychological journal article faster and more accurately than a graduate student, who can do so better than an undergraduate. One is frequently surprised, and often impressed, by how much one

has inferred from what one has heard or read. There has been a great deal of progress in understanding the nature of the skills that expert chess players exhibit, with near-consensus that its chief component is enormous quantities of practice-based knowledge (see Charness, 1991; Ericsson & Smith, 1991). For example, because chess masters tend to remember possible positions much better than random arrangements of pieces while novices do not, we have come to believe that chess masters have stored great numbers of experienced patterns or schemas for their encoding. What LSA would add is that judged similarity between positions should be predictable from a correct dimensionality SVD of a player's history of games, that is, from a matrix of positions by played, watched, or studied games. There is evidence that advanced chess expertise is most consistently acquired from voluminous study of past games, and that its principal skill component is the generation of desirable next moves. Quite possibly LSA could simulate chess experts' judgments of position similarity, thus of likely next moves, by analysis of a body of recorded chess games. Conceivably proficient play could even be generated by choosing from allowable moves, using a few plies of forward evaluation, those most similar to positions from winning games and least similar to those from losers. Perhaps such similarity relations stand behind an expert's poorly articulatable intuitions (in the sense that expert verbalizations may tell less able players little that they can use effectively) about the value of a move or board position.

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 representation is fine. For words that generate a few closely related meanings, or one highly dominant meaning, it is a good compromise. This is the case for the vast majority of word types, but, unfortunately, not for the vast majority of word tokens, because relatively frequent words like line and fly and bear tend to have more senses, as this phenomenon is traditionally described, and senses that are more equally used than do infrequent words. For words that are seriously ambiguous when standing alone, such

as line, ones that might be involved in two or more different meanings with nearly equal frequency, this would appear to be a serious flaw. The average LSA vector for a balanced homograph like bear can bear little similarity to either of its two major meanings. However, we see later that while this raises an issue in need of resolution, it does not necessarily prevent LSA from simulating contextual meaning.

It seems manifest that skilled readers disambiguate words as they go. The introspective experience resembles that of perceiving an ambiguous figure; one or another interpretation quickly becomes dominant and others are lost to awareness. Lexical priming studies beginning with Ratcliff and McKoon (1978) and Swinney (1979) suggest that ambiguous words first activate multiple interpretations, then settle to that sense most appropriate to their discourse context. A dynamic contextual disambiguation process can be mimicked using LSA, but the acquisition and representation of multiple meanings of single words cannot. Consider the sentence, The player caught the high fly to left field. Based on the encyclopedia-based word space, the vector average (the multidimensional mean, or centroid) of the words in this sentence has a cosine of 0.37 with ball , 0.31 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 0.17, 0.03, 0.18, and 0.13 with insect, zipper, 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 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 and is nearly orthogonal to any of the other words, is useless for establishing the topical focus of the discourse. More extensive simulations of LSA-based contextual disambiguation, and their correlations with empirical data on

text comprehension will be described later. Meanwhile, recall the discussion above in which the question was raised as to whether words have multiple stored sense representations or whether different interpretations are only generated dynamically.

Context-based techniques for lexical disambiguation have been tried in computational linguistic experiments with reasonably good results (Grefenstette, 1994; Schutze, 1992; 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 sense representatives be added to an LSA-based representation? As discussed earlier, one hypothetical way to take this step for LSA would be to add initially empty nodes to represent senses, then also find both an analysis architecture that would result in the extra nodes acquiring sense-specific representations and a dynamic performance model that would effect the on-line disambiguation. Such a development is beyond the current implementation of the model. Nonetheless, a sketch of how it might be accomplished is illuminating and will set the stage for later conjectures and questions with regard to text comprehension.

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. Suppose that the input for LSA were a three-way rather than a two-way matrix, with rows of words, columns of paragraphs, and ranks of phrases. Like paragraphs, the phrases would never, or hardly ever, repeat. Cells would contain the (transformed) frequency of co-occurrence of a word, a phrase and a paragraph. (A neural network equivalent might have an additional layer of nodes. Note that in either case, the number of such nodes will be enormous, the computational barrier referred to earlier. Presumably, the brain, using its hundreds of billions of mostly parallel computational elements is not similarly limited in its corresponding process.)

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, any phrase occurring in any paragraph, and so forth, 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 the dynamic performance model might start with the centroid 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.

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 combination of their words. By hypothesis, the various unintended meaning components of the many different words in a passage will tend to be unrelated, to point in many directions in meaning hyperspace, while their vector average will reflect 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.

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 et al., 1993), 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 referred to 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 centroid of the vectors of the words it contained. Finally, the coherence of each paragraph was re-estimated 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 measure comprehensibility by correlating surface structure word types in common between successive sentences (i.e., computing cosines between vectors in the full-dimension 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.

Insert Figure 5 about here

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 homograph or to words not contained in the passages but related inferentially to

each graphemic word type is retrieved—the customary average vector for each word. Following this, there are two possibilities, depending on whether one assumes single or multiple representations for words. Assuming only a single, average representation for each word, the next step (2) is computation of the vector average (centroid) 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, decay. 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. suggest the latter interpretation in that inappropriate homograph relatives lose priming faster than inference words acquire it, but there may be other explanations for this result. In any event, the current LSA representation can only simulate the meaning relations between the words and passages and is indifferent to which of these alternatives, or some other, is involved in the dynamics of comprehension.

In particular, LSA predicts that (1) there should be larger cosines between the "homograph" word and both of its related words than between it and control words; (2) the centroid of the passage words coming before the "homograph" word should have a higher cosine with the context-relevant word related to it than to the context-irrelevant word; and (3) the centroid 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. 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

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. balanced the words by switching words and paragraphs with different meanings and included equal numbers of nonwords. In three experiments of this kind, the principal findings were: (1) in agreement with Ratcliff and McKoon (1978) and Swinney (1979), both senses of an ambiguous word were primed immediately after presentation; (2) by about 300 ms later only the context appropriate associates remained significantly primed; (3) words related to inferred situational themes were not primed at short intervals, but were at delays approaching one second.

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. experiment might go like this: (1) First, a central meaning for

or through the critical homograph were computed as the vector average of the words. The results are shown in Table 1. Following is a summary of the results:

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).
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 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 centroids 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. list. On average, the passages were significantly closer to the appropriate than to either the inappropriate inferentially related words or unrelated control words (above comment relevant here as well).

Insert Table 1 about here

These word and passage relations are fully consistent with either LSA counterpart of the construction-integration theory as outlined earlier. In particular, they show that an LSA based on 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. experiment given the right activation and interaction dynamics. Because homographs are similar to both tested words presumably related to different meanings, homographs 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 representation 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, & Atchley ,1995; Lund & Burgess, in press) who mimicked other priming data using a high-dimensional semantic model, HAL, that is closely related to LSA.[9] 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 Lund et al. have been able to mimic semantic priming results originally reported by Shelton and Martin (1992) as well as some of their own that contrast pairs derived from free-association norms and pairs with intuitively similar meanings—interpreting their high-dimensional word vectors as representing primarily semantic relations. The principal difference between the HAL and LSA approaches to date is our focus on the importance of

dimension matching as a fundamental inductive mechanism rather than merely a computational convenience. However, differences in the analyses and representations provide additional hints and suggestions regarding the construction of such models and interpretation of their results. For example, as outlined earlier, we believe that the use of corpora of a similar size and content to that from which an individual human would have learned the word knowledge that is tested is important if we wish to evaluate the sufficiency of the posited mechanisms. The Lund et al. sample of 160 million words is at least 10 times as much text as college-age priming study participants would have read. However, priming studies involve a different, possibly more sensitive, measure of similarity from our synonym tests, and, for the most part, involve much more common words. Therefore, the relations tested might well be sensitive to the cumulative effects of exposure to both reading and speech. Thus, for this purpose the larger corpus of more nearly conversational content does not seems ill suited. For example, counting speech at an average rate of 120 words per minute, one would need only assume the added experience of about 3 hours per day of continuous speech to bring the total lexical exposure up from our reading estimates to the Lund et al. corpus size.

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 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 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. LSA might be viewed as providing

such models with a new tool for more objective simulation. 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 the 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, as sketched above, including extensions involving iterative construction of context-dependent superstructures, might present a viable alternative to psycholinguistic models based on more traditional linguistic processes and representations.

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-aged 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, in particular words, 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 the same dimensionality as the source, a receiver can, in principle, 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 many direct and indirect relations among a large number of elements and, therefore, one in which such an induction method might play an important role.

We implemented the dimension-matching induction method as a mathematical matrix decomposition method called singular value decomposition, and tested it by simulating the acquisition of vocabulary knowledge from a large body of text. After analyzing and rerepresenting 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 foreign student who had taken this test for admission to a U.S. college. The model's synonym test performance depended strongly on the dimensionality of the representational space into which it fit the words. It did poorly when it relied only on local co-occurrence (too many dimensions), well when it assumed around 300 dimensions, and poorly again when it tried to represent all its word knowledge in much less than 100 dimensions. From this, we concluded that dimension-matching induction 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 as measured by multiple-choice tests) performance, we simulated the acquisition of vocabulary knowledge. The model simulations learned at a rate—in total vocabulary words added per paragraph read—approaching that of children and considerably exceeding learning rates that have been attained in laboratory attempts to teach children word meanings by context (and measured in the same way). Additional simulations showed that the model, when emulating a late grade-school child, acquired roughly three-fourths 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 test words, and estimated that three-fourths as many

total vocabulary words would go from incorrect to correct per paragraph read in the later case as in the former.

Given that the input to the model was data only on co-occurrence of words in passages, so that LSA had no access to word-similarity information based on syntax, logic, or perceptual world-knowledge, all of which can reasonably be assumed to be additional evidence that a dimension matching 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 we 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 processing. Finally, we reported some reanalyses of 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.

At this juncture, we believe the dimension-matching method offers a promising solution to the ancient puzzle of human knowledge induction. It still remains to be determined how wide its scope is among human learning and cognition phenomena—is the model 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, 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 the

mind would choose the optimal dimensionality for its representations, whether by biology or computation. And it remains to be explored 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 syntactically based knowledge, or, perhaps, syntax itself. On the basis of the empirical results and conceptual insights that the LSA theory has already provided we believe that such explorations are worth pursuing.

References

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

Berry, M. W. (1992). Large scale singular value computations. International Journal of Supercomputer Applications, 6(1), 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 and Bradford Books.

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

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

Charness, N. (1991). Expertise in chess: The balance between knowledge and search. In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise. Cambridge, England: Cambridge University Press.

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

Christie, A. (1942). The moving finger. London: Dodd, Mead.

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: Lawrence Erlbaum.

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

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(6), 391–407.

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: Lawrence Erlbaum.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), National Institute of Standards and Technology text retrieval conference (NIST special publication). Washington, DC: NIST.

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 & Bacon.

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

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise (pp. 1–38). Cambridge, England: Cambridge University Press.

Ericsson, A., & Kintsch, W. (1995). Long-term working memory. Psychological Review, 102, 211–245.

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 key-word information systems. The Bell System Technical Journal, 62, 1753–1804.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.

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

Gernsbacher, M. A. (1990). Language comprehension as structure building. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Grolier Academic American Encyclopedia (CD-ROM version). (1990). Danbury, CT: Grolier Electronic Publishing.

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

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. American Educational Research Journal, 21(4), 767–787.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. A. Church (Eds.), Punishment. New York: Appleton.

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 (Eds.), Perspectives of memory research (pp. 325–366). Hillsdale, NJ: Lawrence Erlbaum.

Landauer, T. K. (1986). How much do people remember: Some estimates of the quantity of learned information in long-term memory. Cognitive Science, 10(4), 477–493.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore & J. F. Lehman (Eds.), Cognitive science society (pp. 660–665). Hillsdale, NJ: Lawrence Erlbaum.

Lund, K., & Burgess, C. (In press). Hyperspace analog to language (HAL): A general model of semantic representation (abstract). Brain & Cognition.

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (Under review). Text coherence, background knowledge, and levels of understanding in learning from text.

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

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

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

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.

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: Lawrence 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. (1994). The language instinct: How the mind creates language. New York: William Morrow.

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

Quine, W. V. O. (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. 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.

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. New York: Appleton-Century-Crofts.

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

Seashore, R. H. (1947). How many words do children know? The Packet, 2, 3–17.

Schutze, H. (1992). Context space. In Fall Symposium on probability and natural language. Cambridge, MA.: American Association for Artificial Intelligence.

Shepard, R. N. (1987). Toward 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: Lawrence 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(2), 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 Reviews, 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: Lawrence 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 (Ed.), Analyzing languages in restricted domains: Sublanguage description and processing. Hillsdale, N.J.: Lawrence Erlbaum.

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

Appendix

An Introduction to Singular Value Decomposition and an LSA Example

A well known proof in matrix algebra asserts that any rectangular matrix is equal to the product of three other matrices of a particular form. One of these 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 basis vectors, factors, or dimensions. In SVD they are called singular vectors. The second matrix has columns corresponding to the original columns, but m rows composed of derived singular vectors. The third matrix is a so-called diagonal matrix; that is, it is a square m by 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 labeled the rows and columns terms and contexts. The legend under the diagram defines SVD more formally.

Insert Figure A1 about here

The fundamental proof of SVD shows that there always exists a decomposition of this form such that multiplication of the three derived matrices will reproduce exactly the original matrix 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 LSA, if one or more factor is omitted (that is, 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 removing those with the smallest effect on the error variance 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 will not be described here. Suffice it to say that cookbook versions of them adequate for small (e.g., 1,000 x 1,000) matrices are available in several places (e.g., Mathematica) and a free software version (Berry, 1992) suitable for large matrices such as the one used here to analyze an encyclopedia can currently be obtained from _____. With this algorithm and a high-end work-station with ca 100 M of main memory, matrices on the order of 50,000 by 50,000 (e.g., 50,000 stimuli and 50,000 contexts) can currently be decomposed into representations in up to 400 dimensions with a few hours of computation. The computational complexity is _____. A rough rule of thumb for processing time and storage capacity requirements is _____. Thus, while 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.

Here is a small example of LSA/SVD that gives the flavor of the analysis and demonstrates some of what it accomplishes. This example uses as contexts just the titles of nine technical articles, five about human–computer interaction, four about mathematical graph theory. The original matrix has nine columns, and we have given it 12 rows, each row corresponding to a content word occurring in at least two contexts. (In LSA analyses of text, including some of those reported above, we often omit words that appear in only one sample in doing the SVD. These contribute little to derivation of the space, their vectors can be constructed after the SVD with little loss as the centroid of words in the sample in which they occurred, and their omission sometimes greatly reduces the computation; see Deerwester et al., 1990, and Dumais, 1994, for more on such

details.) The example contexts, with the extracted terms italicized, and the corresponding word-by-context matrix are shown in Figure A2. The complete Singular Value Decomposition of this matrix in nine dimensions is shown in Figure A3. Its cross multiplication would perfectly (ignoring rounding errors) reconstruct the original . Finally, a reduced dimensionality representation, one using only the two largest dimensions, and the reconstruction it generates, are shown in Figure A4. This figure shows a reduction to just two dimensions that approximates the original matrix.

Insert figures A2, A3, and A4 about here

Very roughly and anthropomorphically, SVD, with only values along two orthogonal dimensions to go on, must 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 1 and so much of much of abstract concept 2, and this word has so much of concept 1 and so much of concept 2, and combining those two pieces of information (by linear vector arithmetic), my best guess is that word X actually appeared 0.6 times in document Y."

Comparing the rows for the words human and user in the original and in the two dimensionally reconstructed matrices (Figure A4) shows that while they were totally uncorrelated in the original—the two words never appeared in the same context—they are quite strongly correlated (r = .9) in the reconstructed approximation. Thus, SVD has done just what is wanted. When the contexts contain appropriate "concepts," SVD has filled them in with partial values for words that might well have been used but weren't.

The boxed cell entries in the two tables show this phenomenon in a slightly different way. The word tree did not appear in graph theory title m4. But because m4 did contain graph and minor the zero entry for tree has been replaced with 0.66, which can be viewed as an estimate of how many times it would occur in each of an infinite sample of contexts containing graph and minor. By contrast, the value 1.00 for survey, which appeared once in m4, has been replaced by 0.42, which

reflects the fact that it is unexpected in this context and should be counted as unimportant in characterizing the context itself. Notice that if we were to change the entry in any one cell of the original, the values in the reconstruction with reduced dimensions might be changed everywhere.

Author Note

We gratefully acknowledge the valuable collaboration of Karen Lochbaum of US West Advanced Technologies in the analysis of the Till, Mross, and Kintsch data, and we thank Peter Foltz and Walter Kintsch for many useful discussions and for allowing us to report results data from a joint unpublished study.

Footnotes

[1] For simplicity of exposition, we are intentionally imprecise here in the use of the terms distance and similarity. In the actual modeling, similarity is measured as 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 embedded. Thus, at least at a qualitative level, the two vocabularies for describing the relations are equivalent.

[2] Although this exploratory process will take some advantage of chance, there is no reason why any choice of dimension should be much better than any other unless some mechanism like the one proposed is at work. The model's remaining parameters were fitted only to its input data and not to the criterion test.

[3] Strictly speaking, the entropy operation is global, added up over all occurrences of the event type (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 US has its effect by appropriately weakening the local representation of the CS–US connection.

[4] We have used cosine similarities because they usually work best in the information retrieval application. It has never been clear why. They 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 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. To stretch speculation, it may also be noted that the excitation of one neuron by another is proportional to the dot product (the numerator of a cosine) of the output of one and the sensitivities of the other across the synaptic connections that they share.

[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 & Hanks, 1990). Because of the transform, this poor result is still better than is 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.

[6] 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 centroid 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.

[7] To estimate the number of words that the learner would see for the first time in a paragraph, we used the log-normal model proposed by Carroll 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 we did use a function of the same form to interpolate the center values used to stand for frequency bands.

[8] 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 .75809. If such a word is met in the next sample—which we call a direct effect—it will have been met three times, and there will have been 25,001 total samples; its z will increases to .83202. By the maximum of three from a normal distribution criterion, its probability of being correct on the TOEFL test will rise by .029461. But the probability of a given word in a sample being a word of frequency two in the corpus is 11,915 * 2)/ 5 * 10 ^ 6 = .0047, so the direct gain in probability correct for a single word actually encountered attributable to words of frequency two is just .000138. However, there is also an indirect gain expected for frequency-two word types that were not encountered—which we call an indirect effect. Adding an additional paragraph makes these words add no occurrences but go from 25,000 samples to 25,001 samples. By equation 1, the z for such a word type goes, on average, from .75809 to .75810, and its estimated probability correct goes up by 2 * 10 ^ -6. But, because there are 11,195 word types of frequency two, the total indirect vocabulary gain is .079120. Finally, we cumulated these effects over all 37 word-frequency bands.

[9] Indeed, there is a direct line of descent between LSA and the HAL model of Burgess and colleagues. Lund et al. (1995) credit an unpublished paper of H. Schutze as the inspiration for their method of deriving semantic distance from large corpora, and Shutze, in the same and other articles (e.g., 1992), cites Deerwester et al. (1990), the initial presentation of the LSA method for information retrieval.

Table 1

LSA Cosines Between Homographic Words and Context Passages

Target "Prime"	A. Appropriate sense	B. Inappropriate sense	C. Appropriate inferred	D. Inappropriate inferred	E. Unrelated control
Homograph alone	.20	.21	.09	.05	.07 p vs. a or b < 10^{-5} z = .84
Full passage with homograph	.24	.21	.21	.14 p vs. c = .0008 z = .62	.15 p vs. c = .0004 z = .56
Full passage without homograph	.21	.15 p vs. a = .006 z = .49	.20	.14 p vs. c = .004 z = .51	.16 p vs. c = .006 z = .42

Note. From Till, Mross, & Kintsch (1988), Experiment 1.

Figure Captions

Figure 1. Added value of knowing the correct dimensionality for estimating interpoint distances.

Figure 2. Schematic illustration of dimension reduction in Singular Value Decomposition. The corresponding gray rows are original 30,000 and condensed 300 dimensional vectors for the same word.

Figure 3. Effect of number of dimensions in LSA representation on synonym test performance.

Figure 4. Discrimination rations, z, between cosines for correct and incorrect alternatives on the TOEFL synonym test as a function of total number of text segments analyzed by LSA and, as the parameter, the mean number of text segments containing a word that was tested.

Figure 5. LSA and textual coherence. Correspondence between LSA and lexical overlap estimates of conceptual similarity of succeeding sentences and the comprehensibility of paragraphs.

Figure A1. The Singular Value Decomposition (SVD) of a rectangular term by context matrix, X, where: T has orthogonal unit-length columns (T' T = I); C has orthogonal unit-length columns (C' C = I); S is the diagonal matrix of singular values; t is the number of rows of X; c is the number of columns of X; and m is the rank of X {< = min (t, c)}.

Figure A2. Nine text context are shown at the top, five titles of papers about human–computer interaction, four about mathematical graph theory. A matrix of contexts by words contained in them is shown below. Only words occurring in at least two contexts were retained. Cells contain the number of times a particular word occurs in a given context. (In LSA the cell entries are ordinarily transformed to ln x/ entrophy (x); here we left them raw for clarity of exposition.)

Figure A3. The full dimensional Singular Value Decomposition of the matrix of Figure A2.

Figure A4. The two dimensional SVD of the matrix in Figure A2, and the reconstruction that multiplication of the three component matrices generates. The grayed rows and cells illustrate how the dimension reduction has induced similarity of meaning (see text).

University of Colorado at Boulder

Institute of Cognitive Science

Muenzinger Psychology Building
Campus Box 344
Boulder, Colorado 80309-0344
(303) 492-5063
FAX: (303) 492-7177
email: cogsci@psych.colorado.edu
URL: http://psychwww.colorado.edu/ics/