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 <u>either</u> always precede nouns, <u>or</u> 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 <u>some</u> 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