1

LSA as a Theory of Meaning

Thomas K Landauer
Pearson Knowledge Technologies and University of Colorado

The fundamental scientific puzzle addressed by the latent semantic analysis (LSA) theory is that there are hundreds of distinctly different human languages, every one with tens of thousands of words. The ability to understand the meanings of utterances composed of these words must be acquired by virtually every human who grows up surrounded by language. There must, therefore, be some humanly shared method—some computational system—by which any human mind can learn to do this for any language by extensive immersion, and without being explicitly taught definitions or rules for any significant number of words.

Most past and still popular discussions of the problem focus on debates concerning how much of this capability is innate and how much learned (Chomsky, 1991b) or what abstract architectures of cognition might support it—such as whether it rests on association (Skinner, 1957) or requires a theory of mind (Bloom, 2000).

The issue with which LSA is concerned is different. LSA theory addresses the problem of exactly how word and passage meaning can be constructed from experience with language, that is, by what mechanisms—instinctive, learned, or both—this can be accomplished.

Carefully describing and analyzing the phenomenon has been the center of attention for experimental psychology, linguistics, and philosophy. Other areas of interest include pinpointing what parts of the brain are most

heavily involved in which functions and how they interact, or positing functional modules and system models. But, although necessary or useful, these approaches do not solve the problem of how it is possible to make the brain, or any other system, acquire the needed abilities at their natural scale and rate.

This leads us to ask the question: Suppose we have available a corpus of data approximating the mass of intrinsic and extrinsic language-relevant experience that a human encounters, a computer with power that could match that of the human brain, and a sufficiently clever learning algorithm and data storage method. Could it learn the meanings of all the words in any language it was given?

The keystone discovery for LSA was that using just a single simple constraint on the structure of verbal meaning, and a rough approximation to the same experience as humans, LSA can perform many meaning-based cognitive tasks as well as humans.

That this provides a proof that LSA creates meaning is a proposition that manifestly requires defense. Therefore, instead of starting with explication of the workings of the model itself, the chapter first presents arguments in favor of that proposition. The arguments rest on descriptions of what LSA achieves and how its main counterarguments can be discounted.

THE TRADITIONAL ANTILEARNING ARGUMENT

Many well-known thinkers—Plato, Bickerton (1995), Chomsky (1991b), Fodor (1987), Gleitman (1990), Gold (1967), Jackendoff (1992), Osherson, Stob, and Weinstein (1984), Pinker (1994), to name a few—have considered this *prima facie* impossible, usually on the grounds that humans learn language too easily, that they are exposed to too little evidence, correction, or instruction to make all the conceptual distinctions and generalizations that natural languages demand. This argument has been applied mainly to the learning of grammar, but has been asserted with almost equal conviction to apply to the learning of word meanings as well, most famously by Plato, Chomsky, and Pinker. Given this postulate, it follows that the mind (brain, or any equivalent computational system) must be equipped with other sources of conceptual and linguistic knowledge. This is not an entirely unreasonable hypothesis. After all, the vast majority of living things come equipped with or can develop complex and important behavioral capabilities in isolation from other living things. Given this widely accepted assumption, it would obviously be impossible for a computer using input only from a sample of natural language in the form of unmodified text to come even close to doing things with verbal meaning that humans do.

THE LSA BREAKTHROUGH

It was thus a major surprise to discover that a conceptually simple algorithm applied to bodies of ordinary text could learn to match literate humans on tasks that if done by people would be assumed to imply understanding of the meaning of words and passages. The model that first accomplished this feat was LSA.

LSA is a computational model that does many humanlike things with language. The following are but a few: After autonomous learning from a large body of representative text, it scores well into the high school student range on a standardized multiple-choice vocabulary test; used alone to rate the adequacy of content of expository essays (other variables are added in full- scale grading systems; Landauer, Laham, & Foltz, 2003a, 2003b), estimated in more than one way, it shares 85%–90% as much information with expert human readers as two human readers share with each other (Landauer, 2002a); it has measured the effect on comprehension of paragraph-to-paragraph coherence better than human coding (Foltz, Kintsch, & Landauer, 1998); it has successfully modeled several laboratory findings in cognitive psychology (Howard, Addis, Jing, & Kahana, chap. 7 in this volume; Landauer, 2002a; Landauer & Dumais, 1997; Lund, Burgess, & Atchley, 1995); it detects improvements in student knowledge from before to after reading as well as human judges (Rehder et al., 1998; Wolfe et al., 1998); it can diagnose schizophrenia from what patients say as well as experienced psychiatrists (Elvevåg, Foltz, Weinberger, & Goldberg, 2005); it improves information retrieval by up to 30% by being able to match queries to documents of the same meaning when there are few or no words in common and reject those with many when irrelevant (Dumais, 1991), and can do the same for queries in one language matching documents in another where no words are alike (Dumais, Landauer, & Littman, 1996); it does its basic functions of correctly simulating human judgments of meaning similarity between paragraphs without modification by the same algorithm in every language to which it has been applied, examples of which include Arabic, Hindi, and Chinese in their native orthographic or ideographic form; and when sets of all LSA similarities among words for perceptual entities such as kinds of objects (e.g., flowers, trees, birds, chairs, or colors) are subjected to multidimensional scaling, the resulting structures match those based on human similarity judgments quite well in many cases, moderately well in others (Laham, 1997, 2000), just as we would expect (and later explain) because text lacks eyes, ears, and fingers.

I view these and its several other successful simulations (see Landauer, 2002a; Landauer, Foltz, & Laham, 1998) as evidence that LSA and models like it (Griffiths & Steyvers, 2003; Steyvers & Griffiths, chap. 21 in this vol-

ume) are candidate mechanisms to explain much of how verbal meaning might be learned and used by the human mind.

ABOUT LSA'S KIND OF THEORY

LSA offers a very different kind of account of verbal meaning from any that went before, including centuries of theories from philosophy, linguistics, and psychology. Its only real predecessor is an explanation inherent in connectionist models but unrealized yet at scale (O'Reilly & Munakata, 2000). Previous accounts had all been in the form of rules, descriptions, or variables (parts of speech, grammars, etc.) that could only be applied by human intercession, products of the very process that needs explanation. By contrast, at least in programmatic goal, the LSA account demands that the only data allowed the theory and its computational instantiations be those to which natural human language users have access. The theory must operate on the data by means that can be expressed with mathematical rigor, not through the intervention of human judgments. This disallows any linguistic rule or structure unless it can be proved that all human minds do equivalent things without explicit instruction from other speakers, the long unattained goal of the search for a universal grammar. It also rules out as explanations—as contrasted with explorations—computational linguistic systems that are trained on corpora that have been annotated by human speakers in ways that only human speakers can.

This way of explaining language and its meaning is so at odds with most traditional views and speculations that, in Piaget's terminology, it is hard for many people, both lay and scholar, to accommodate. Thus, before introducing its history and more of its evidence and uses, I want to arm readers with a basic understanding of what LSA is and how it illuminates what verbal meaning might be.

BUT WHAT IS MEANING?

First, however, let us take head-on the question of what it signifies to call something a theory of meaning. For a start, I take it that meaning as carried by words and word strings is what allows modern humans to engage in verbal thought and rich interpersonal communication. But this, of course, still begs the question of what meaning itself is.

Philosophers, linguists, humanists, novelists, poets, and theologians have used the word "meaning" in a plethora of ways, ranging, for example, from the truth of matters to intrinsic properties of objects and happenings in the world, to mental constructions of the outside world, to physically irreducible mystical essences, as in Plato's ideas, to symbols in an internal communication and reasoning system, to potentially true but too vague no-

tions such as how words are used (Wittgenstein, 1953). Some assert that meanings are abstract concepts or properties of the world that exist prior to and independently of any language-dependent representation. This leads to assertions that by nature or definition computers cannot create meaning from data; meaning must exist first. Therefore, what a computer creates, stores, and uses cannot, *ipso facto*, be meaning itself.

A sort of corollary of this postulate is that what we commonly think of as the meaning of a word has to be derived from, "grounded in," already meaningful primitives in perception or action (Barsalou, 1999; Glenberg & Robertson, 2000; Harnad, 1990; Searle, 1982). In our view ("our" meaning proponents of LSA-like theories), however, what goes on in the mind (and, by identity, the brain) in direct visual or auditory, or any other perception, is fundamentally the same as what goes on in any other form of cognition and has no necessary priority over other sources of knowledge, such as—in particular—autonomous manipulations of strings of words that convey abstract combinations of ideas such as imaginary numbers. Of course, strings of words must somehow be able to represent and convey both veridical and hypothetical information about our inner and outer worlds; otherwise, language would not be very useful. Certainly, that is, much perceptual experience must map onto linguistic expressions. And many linguistic expressions must map onto perceptual experience. However, once the mappings have been obtained through the cultural evolution of a language, there is no necessity that most of the knowledge of meaning cannot be learned from exposure to language itself. The highly developed verbal-intellectual feats of Helen Keller, and the more modest but still near normal knowledge and communication accomplishments of most congenitally blind people—including the correct use of color and shape words—would be impossible (Keller, 1905; Landau & Gleitman, 1985).

This puts the causal situation in a different light. We may often first learn relations of most words and passages to each other from our matrices of verbal experiences and then attach them to perceptual experience by embedding them in the abstract word space. Take the example of geographical maps. A map of England's cities can be constructed from a relatively small set of measured point-to-point distances projected onto the surface of a sphere. You can understand the geography of England simply by viewing the map. I can tell you that Cambridge is North of London and Oxford north of Cambridge, and you can then tell me that Oxford is north of Cambridge (from the map, not the logic).

It is important to understand that in LSA, as in a map, the coordinates are arbitrary. North and south are conventionally used for the earth, but the relation of any point to any other would be just as well located by any other set of nonidentical axes. LSA axes are not derived from human verbal descriptions; they are underlying points in a coordinate system, in LSA's case,

one that relates meanings to each other. LSA's theory of meaning is that the underlying map is the primitive substrate that gives words meaning, not vice versa. By contrast, artificial intelligence (AI) ontologies, such as WordNet and CYC, start with intuitive human judgments about relations among words, the output of the mechanism LSA seeks to provide.

In LSA, words do not have meanings on their own that define the axes, words get their meanings from their mapping. Nonetheless, it is sometimes possible to rotate the space so that at least some words, as discrete points, fall near common, not necessarily orthogonal, axes so that word names can be associated with them to yield intuitive interpretation. Some other LSA-like systems have been built to maximize such intuitiveness (Griffiths & Styvers, 2003).

ON THE EPISTEMOLOGICAL NATURE OF LSA

Now a map is not the thing itself; it is an abstraction from which much more can be induced—an infinite number of point-to-point distances computed by triangulation from earlier established points—than is possible with only raw perceptual experiences of the real thing, say from walking around England's green and pleasant land. Just so, language maps perceptions and actions onto the physical world, and vice versa, but does very much more by supporting induction of an infinite number of meanings of words and word combinations. It is, according to LSA, almost entirely the relations that are represented and activated by words and collections of words that create verbal meaning. And it is primarily these abstract relations that make thinking, reasoning, and interpersonal communication possible. Qualitatively, this proposal shares much with the ideas of Wittgenstein (1953), but as we will see, LSA transforms them into a concrete and testable mathematical theory.

However, there is another noteworthy difference between many abstract theories and LSA. The difference concerns the unique nature of the phenomenon with which LSA deals. Memory and language are not physical objects, they are properties of an information-processing system. Their nature is only present in information storage, organization, and control. Thus, LSA is not only a mapping technique that is not the real thing—a computer, not a brain—it is a real thing in the same sense as thought is a real thing. It not only models, it does some of the same things. In this way, it is a bit unusual as a model. Bohr's model of the atom is a marvel of physical explanation, but it cannot actually build physical molecules. Model airplanes or ships can be faithful representations, even fly or sail, but they cannot transport people or cargo. Even most mathematical models in psychology have the same limitation, some neural nets being exceptions.

OTHER MEANINGS

Word meanings are not the only form of meaning. Complex relations among perceptions and actions must be entities of a highly similar sort, the kind shared by nonverbal and preverbal animals and infants. And these "primitive" meanings must also have learned interrelations with verbal meaning that places at least some on the same cognitive map as words. Integrating perceptions into the map also changes the meanings of words and passages. This is an almost self-evident necessity that is fully consistent with our claim that most relations among verbal meanings can be learned from language experience alone. As we shall see, the success of LSA is incontrovertible evidence of this. Our later description of cross-language retrieval by LSA also suggests a mechanism by which the mapping between perception and language might be constructed.

Just as creation of geographical maps from a small number of observations allows induction of greatly more relations, if the meaning of verbal expressions is a structure of the same sort, then most word–word and word–perception relations should be inducible from measures of a small subset of such relations. The question of how this is done can be approached by finding computable models that can accomplish the same thing. LSA is one such model.

LSA IS NOT A COMPLETE MODEL OF LANGUAGE

Lest the scope of our argument be misunderstood, let me make it clear before going on that LSA is not a complete theory of language or meaning. It does not take into account word order by which the meaning of sentences or the implications of sentence and paragraph order are altered. Without human help, it often does not adequately represent the variability of meanings conveyed by predication, anaphora, metaphor, modification, attachment, quantification, logical or mathematical propositions, or negations. These are important issues in language understanding that have not yet been reduced to the kinds of explanation we desire. This fact, however, does not mean that the theory is wrong, only that it does not cover all aspects of language. The analogy of accounting for the trajectory of a falling leaf comes to mind. Thus, contra Popper, constructing examples of sentence-to-sentence similarities for which a model does not match human judgments well (Glenberg & Robertson, 2000) does not falsify the theory. Nor does it show that the general approach to language modeling will not succeed. Indeed, new approaches to modeling the modification of meaning by word order may not be too long in coming. A good start on the latter problem is represented by Dennis's SP model (chap. 3 in this volume, 2005).

All this aside, however, estimates of the relative amount that word order and word choice contribute to overall meaning of a sentence or paragraph suggest that the latter carries the lion's share, on the order of 80%–90% (Landauer, 2002a, and later.)

ABOUT INTUITIVE REVULSION TO LSA

Undoubtedly, however, this approach to the question will fail to satisfy many from the other camps. They will still feel that, at best, computers can only artificially mimic or mirror the real thing, not actually be it themselves. We have no strong quarrel with such a position. LSA is a theory—not reality itself—at least in so far as the particular mathematics it uses are unlikely to be the same. Nonetheless, the fact that LSA can do many of the things that humans do almost indistinguishably from humans means that it must in part be isomorphic to the real thing—whatever it is. Thus, we believe that a successful computational model, even if incomplete, supplies a better foundation for progress in explaining the phenomena of language and meaning than do purely verbal philosophical arguments from which simulations of human performance cannot be constructed without contributions from the knowledge of language that is to be explained.

In any event, however, such arguments do not overly concern us. LSA as a theory of meaning has aimed primarily at a more restricted, empirical, and pragmatic domain for explaining the nature of meaning. We are interested in how to get a machine to do useful things with language even if not the same things in exactly the same way. The question is how experience can be used to learn how to use words and strings of words to do what humans do. It is because LSA can do so that we think it provides a promising theory about how language works. Because by any sensible interpretation, adequate use of words requires knowledge of verbal meaning and LSA makes adequate use of words, LSA must have such knowledge too.

A REVIEW TO HERE

Let me review what I have argued so far. LSA demonstrates a computational method by which a major component of language learning and use can be achieved. It is in that sense that LSA is a theory. It is specifically a theory of meaning because it offers an explanation of phenomena that are ordinarily considered to be manifestations of meaning—the expression, comprehension, and communication of ideas and knowledge in words and passages of words. It offers an explicit theory about the nature of word and passage meaning and its acquisition and application. It makes possible

computer systems that accomplish a wide range of cognitive tasks performed by humans, and often does them essentially as well. This makes its basic mechanism, or something much like it, a candidate for explaining the corresponding human abilities. LSA is a theory about an essential aspect of language, not of everything about language. However, its successes encourage hope that more complete theories in the same spirit, say with additional cooperating mechanisms based on instinct or learning, are possible.

MORE ABOUT THE MISSING PIECES OF THE PUZZLE

However, consider somewhat further the matter of how word order and grammar affect meaning—important influences that unsupplemented LSA ignores. For simplicity, I will sometimes lump together under the term *syntax*, many of the ways in which differences in word order are involved in linguistic descriptions of its effects on meaning: including word class requirements, constituent combination, and higher order sentence structures (see Wikipedia). Note, however, that not all the ways in which grammar and syntax work are excluded by LSA. The combinations of words that best produce a passage meaning must have the right tenses, number, determiners, and so forth. It is only those differences that require differential word order that are at stake.

Given the long and almost exclusive concentration of linguistic research and theory on these factors, how can LSA do so well without them? People in over 2,000 cultures have learned hundreds of easily distinguishable languages, any one of which is almost incomprehensible to speakers of any of the others. Some scholars (Bickerton, 1995; Chomsky, 1991a, 1991b; Fodor, 1987; Gold, 1967; Pinker, 1994) think that the heart of this mystery is the wide variety of grammars governing how different classes of words and the ordering of words in an utterance give rise to differences in meaning (the meaning of words usually treated as a primitive).

This is an important unsolved problem despite long and sophisticated attempts. The principal attack has been to search for an innately given skeleton that is easily transmuted into that for a given language by exposure to a small sample of its use, somewhat as songbirds learn their songs. Unfortunately, we do not know of any detailed mechanism by which this could actually work (even in song birds), much less an algorithm that can accomplish the feat. However, we believe that an even more important but tractable problem (as LSA's success without word order suggests) is how people learn the meaning of all the words in their language. English is probably the champion, having many millions of sometimes-used words of which a well-educated adult must have reasonable command of the meaning of around 100,000. Following Plato (and some others in between), Chomsky

(1991a, 1991b) has averred that this is simply impossible because the exposure to words in informative contexts is manifestly much too limited. Thus, he (and others) have concluded that infants must start out with all the possible concepts needed in any human environment and simply learn names to go with a large number of them. Thus, the acquisition of meaning for words is finessed, the meanings preexist in all of us—supposedly learning names for all of concepts takes very little new knowledge, just as picking the right grammar supposedly does. In both cases, I find such claims unsatisfactory, given the difference, say, between the concepts needed by a Chinese rice farmer, a French biochemist, and an Inuit seal hunter, and between the strongly interrelated use of grammars to combine them into meaningful utterances. But that is not my major complaint about the hypothesis. The most unacceptable part is the basis of the hypothesis, which in a nutshell can be expressed as such: "I cannot imagine any way that all the concepts and words could be learned with the evidence available." Before a problem is solved, people often cannot imagine a solution. That humans have a species special instinctive capacity for language, as argued by Pinker (1994) and others (e.g., Ridley, 1994), is unexceptionable, but does not answer the more interesting question of how the instinct works.

Thus, I do not argue that humans have no innate knowledge relevant to language. Without experience, centipedes and foals can walk, and bees can navigate; surely humans come ready for the cognitive tasks that all humans must perform. From worms to humans, the brain/mind must be extremely flexible, able to adapt dramatically when so required. Learning a particular vocabulary of tens of thousands of words must be one of those situations.

What is needed, then, is a mental mechanism that actually can learn language from the available evidence. Presumably, this must be a previously unknown mechanism because the theories of learning previously available to world-class thinkers like Chomsky (1991a, 1991b), Fodor (1987), Skinner (1957), Pinker (1994), and Pylyshyn (1980), did not suffice. Such a mechanism must instantiate a computation that can do what's needed.

To bring this to an end, let me summarize: LSA provides one way to do very much of what's needed. LSA falls short of what human minds can do in several ways, some quite important. But, the really important thing it does do is provide a computational theory of what word and passage meaning is and how it can be acquired from the evidence available to an ordinary human. The theory is proposed not only as an idealization or abstraction of the actual mechanism, but as a computational information-processing mechanism that actually performs many of the same functions. However, it is not claimed that its computational method is the same as nature's except at a higher level of abstraction that includes it as an example. What is claimed to be the same as human's is the general kind of function computed.

THE COMPUTATION EMPLOYED BY LSA

Finally, assuming that readers are now willing to grant the possibility that LSA qualifies as a theory of meaning—even if they still object to the way we have operationalized the term—we are ready to see what LSA is and how it does what it does. The description will still be a high-level conceptual account, the real math left for Martin and Berry (chap. 2 in this volume) and other sources (Berry, Dumais, & O'Brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

The most important foundation for LSA, especially given the way we have introduced the problem, is the power of exploiting mutual constraints. LSA rests on a single conceptually simple constraint, that the representation of any meaningful passage must be composed as a function of the representations of the words it contains. This constraint, somewhat generalized, goes under the name of "compositionality." The particular compositional constraint imposed by LSA is that representations of passages be sums of its representations of words. Basic algebra gives the most familiar example of how this constraint supplies the inductive power needed. Consider these two simultaneous equations: A + 2B = 8 and A + B = 5.

As all algebra students know, neither equation alone tells the value of either A or B, but the two together tells both. In the very same way, in LSA the meaning of a passage of text is the sum of the meanings of its words. In mathematical form:

$$\text{meaning passage} = \Sigma(m_{\text{word 1,}} m_{\text{word 2,}} \ldots m_{\text{word n}}). \qquad \textbf{1.1}$$

Thus, LSA models a passage as a simple linear equation, and a large corpus of text as a large set of simultaneous equations. (The mathematics and computations, singular value decomposition, by which the system is usually solved, are spelled out by Berry, 1992, and Martin & Berry, chap. 2 in this volume.) The constraint-satisfaction approach is also used by Dennis and by Steyvers and Griffiths (chaps. 3 and 21, respectively, in this volume), but with different constraints. To create an LSA representation of word meanings that satisfies this condition, one first secures a large (ideally, but never completely) representative sample of the language experience of people that is typical in content and size to that experienced by people whose language is to be captured in the model. Ideally, this would include all natural exposures to and uses of language, including their perceptual, physiological, and mind/brain contexts. This being unavailable, we have used the best approximation that we can find and fit into a computer, a large corpus of text that has been sampled so as to represent what a normal human would have read. (Landauer & Dumais, 1997, estimated that up to 80% of the words known to a college freshman would have been met only in print,

and those met in speech would almost all have been met many times in print as well.) This is called a training corpus, which then divides the corpus of text into segments that carry (ideally again) full and coherent meanings, typically paragraphs. From this, one constructs a matrix with a row for each unique word type and a column for each passage; the cells contain (an information-theoretic transform of) the number of times that a particular word type appears in a particular passage.

As should be apparent from this description, the solution for any one word may depend on the solution of many other words; indeed, changing any one could, in principle, change every other. Therefore, successful simulation of human word and passage meaning can depend strongly on giving it a sufficiently large and representative text corpus to learn from, just as humans need vast experience to learn their languages. Hart and Risley (1995) estimated from large-scale systematic observations that an average child hears around six million word tokens per year. In practice, we have usually found that corpora containing from 10^7 to 10^{10} words of text divided into 10^5 to 10^9 paragraphs with 10^5 to 10^6 different word types—amounts of language experience roughly equivalent to that of children to highly literate adults—produce good results.

This produces a set of 10^5 to 10^9 simultaneous linear equations. Leaving out many important details, this system of simultaneous equations is solved for the meaning of each word-type and each passage, with a solution in which every paragraph conforms to the additive equation given earlier. Because the sample of paragraphs is very large and representative, we can be reasonably sure that new meaningful paragraphs will conform to the same function. Success in applications, such as scoring novel essays, confirms this expectation.

The solution is in the form of a set of vectors, one for each word and passage, each vector having typically 200–500 elements—factors or dimensions—in a "semantic space." The meaning of any new passage is computed by vector addition of the word vectors it contains. The similarity of meaning of two words is measured as the cosine (or dot product or Euclidean distance, depending on the application) between the vectors, and the similarity of two passages (of any length) as the same measure on the sum or average of all its contained words. Note carefully the important difference between this process and methods that measure the relative frequency of local co-occurrence to estimate the similarity of meaning of words. For example, Lund & Burgess, 1996.

DIMENSION REDUCTION AND ITS IMPORTANCE

I have glossed over the fact that the vectors representing words and passages usually have 200–500 elements rather than 3 or 10,000. It is actually a

matter of critical importance to the success of LSA and of other similar methods (Erosheva, Fienberg, & Lafferty, 2004; Griffiths & Steyvers, 2003; Steyvers & Griffiths, chap. 21 in this volume). Consider again the mapping of geographical points. Suppose you measured the distances between, say, Oslo, Baghdad, and Sydney, and tried to plot them all on the same straight line. They would not fit together. If you try to do it in two dimensions, it gets much better, but still with gross distortions. Using three dimensions—a globe—you get quite close, ignoring elevations, which would take yet another dimension. Suppose now that you now plotted the same positions in one more dimension, say new measures in feet instead of miles. It would be of limited help, yielding greater accuracy than might be needed. A better solution is to find a dimensionality for which the resolution is optimized for your purposes, for example, for words to make near synonyms such as "car" and "automobile" much but not exactly the same, unrelated concepts such as "philosophy" and "automobile" not at all, and distantly related words such as "football" and "algebra" only slightly.

Here is another kind of explanation. With enough variables, every object is different from any other. For too few, all objects can easily be the same. Consider the vectors **|a m r c|** and **|a m x b|**. They are different by two components. If we drop the last two components, they are identical. If we drop just the last component, they are more nearly the same than initially. The effect is much like squinting just the right amount to make two different faces look the same while still looking like faces. If we want a graded similarity function relating words and passages that varies from little to much in a helpful way, we need an intermediate number of dimensions. (Note that it would not do to simply have different granularity on a single dimension because of the mapping problem described earlier.)

Something analogous happens in LSA. Here we want the amount of resolution in the model to match the resolution in human word and passage meanings. Optimal dimension reduction is a common workhorse in analysis of complex problems in many fields of science and engineering. One way of showing its immense importance in LSA is what happens when the original word by paragraph matrix is reconstructed from the reduced dimensional representation—estimating each now by distances from others. Suppose the initial matrix for a corpus has 500 million cells—each containing the number of times one of 50,000 unique word types appears in a particular one of 100,000 paragraphs. Over 99.9% of the cells will turn out to be empty. This makes the comparison of word or paragraph meanings quite chancy. However, after dimension reduction and reconstruction, every cell will be filled with an estimate that yields a similarity between any paragraph and any other and between any word and any other.

This is an extremely powerful kind of induction. It is what accounts for LSA's advantage over most current methods of information retrieval,

which rely on matching literal words (or words that have been stemmed or lemmatized or to be equivalent to a few others). It is also what accounts for its ability to measure the similarity of two essays that use totally different words, and for all of the other properties of LSA that defy the intuition that learning language from language is impossible.

ABOUT CO-OCCURRENCE AND LSA

Sometimes people misunderstand the mechanism by which LSA represents similarity of word meanings as counting the relative number of times that two words appear in the same sentences or passages. Whereas LSA starts with a kind of co-occurrence, that of words with passages, the analysis produces a result in which the fact that two words appear in the same passage is not what makes them similar. As in all simultaneous equation problems, it is the degree to which they have the same effects on their summed values wherever they occur in meaningful passages that is measured in the result. Indeed, in a study (Landauer, 2002a) of a large random sample of word pairs, the correlation between LSA-measured word pair similarities (cosines and several alternate contingency metrics) and the number of times they appeared in the same passage was only a little higher than that with the number of times they appeared separately in different passages, which, by the common notion of co-occurrence, should make them more different, not more similar. In a way, the result is the opposite of word meaning coming from co-occurrence. LSA learns about the meaning of a word from every meeting with it and from the composition of all the passages in which it does not occur. It is only after learning its meaning by SVD and dimension reduction that its relation to all other words can be computed.

Consider also that the fact that two words that appeared in the same sentence would not be very good evidence that they had the same meaning because there would often be more expressive value in using two different words. On the other hand, it might be evidence that they were related to the same topic, and thus reflecting their choice by the author because the meanings of each helped to add up to the desired meaning of the whole, a different direction of causation for local co-occurrence.

Similarly, sometimes the mechanism of LSA has been attributed to indirect co-occurrences: **A** does not occur in the same passage as **B**, but both occur in some third passage, or more elaborately that local co-occurrences are percolated up a hierarchical or other network structure to connect with other words. This seems a quite unlikely mechanism to me. If direct local co-occurrence is not much more effective than separate occurrences, then indirect chains between words do not look promising. In any event, no such model has had the success of LSA, and LSA does not work that way.

It follows the same kind of arguments that the number or proportion of literal words shared between two passages is not the determinant of their similarity in LSA, as is illustrated later.

EXAMPLES OF LSA PROPERTIES

Here are a few more examples of what LSA accomplishes. Results are stated in cosines (which for vectors are ordered in the same manner as correlations). Cosine values can range between –1 and 1, but in practice rarely go below 0 for word–word, passage–passage, or word–passage similarities. Randomly chosen pairs of words from the same corpus as the example have a mean of about .03 and a standard deviation of about .08.

Next are some phrase and sentence cosine similarities where there are no shared words. These examples are selected, not typical, but of a sort that occurs often enough to make a large difference in the model's ability to simulate human similarity judgments:

"Several doctors operated on a patient"
"The surgery was done by many physicians " (cosine = .66)

"A circle's diameter":
"radius of spheres" (cosine = .55)
"music of the spheres" (cosine = .03)

Next are a few examples of what LSA accomplishes, first some typical word–word cosine similarity measures, presented in Table 1.1. Note that when LSA computes a meaning vector for a whole passage, the identities of the literal words of which it was composed are lost. The computation is a one-way function that cannot be reversed. It is possible to search for a set of words that will capture the gist of the meaning by adding up to near the same total. These are not necessarily the words initially used, and of course word order cannot be recovered at all, although one could usually construct a new syntactically correct passage with a highly similar overall meaning. The situation is analogous to human memory for text; a short time after reading a sentence or paragraph, people remember the gist of what they have read but loose the exact wording, a phenomenon known since Bartlett (1932). LSA's paragraph meanings are a form of gist.

WORD SENSES

As follows from the basic additive linear function of LSA, the vector representation of a unique word type is the average effect that it has on the meaning of paragraphs in which it occurs. (It also results from the effects of paragraphs in which it does not appear, but we will ignore that here for sim-

TABLE 1.1
Typical Word–Word Cosine Similarity Measures

Word Pair	*Cosine*
thing–things	.61
man–woman	.37
husband–wife	.87
sugar–sweet	.42
salt–NaCl	.61
cold–frigid	.44
mouse–mice	.79
doctor–physician	.61
physician–nurse	.76
go–went	.71
go–going	.69
going–gone	.54
should–ought	.51
kind–unkind	.18
upwards–downwards	.17
clockwise–counterclockwise	.85
black–white	.72
she–her	.98
he–him	.93
junk–garbage	.37
sun–moon	.28
sun–earth	.46
moon–earth	.40
Nebraska–Kansas	.87
Nebraska–Florida	.25
Kansas–Florida	.24

plicity.) A word vector thus carries all of its "senses"—all the different ways in which it has been used weighted by the relative magnitude of its effects in each paragraph in which it has occurred. Conversely, all of the contextually determined meanings of a word enter into the meaning of a paragraph. There are no separate representations for separate senses of a word type. The notion of disambiguating a word by sense before interpreting a passage containing it—for which a great deal of effort has been spent in computational

linguistics—has no place in the LSA theory. Instead of prior disambiguation, the context in which it appears determines its contribution to meaning. Paragraph meaning is usually not unnaturally distorted by this mechanism for two reasons. First, the proportion of a word's various merged meanings is equal to the importance it gains from the meanings of all its occurrence and nonoccurrences. Thus, it will convey most strongly the right meaning for just those paragraphs in which it occurs. If one broke a word's meaning into discrete components (which we would not do because LSA treats the meaning as a continuous whole), and multiplied the resulting components by their frequency over the whole corpus, then the overall amount of conflict between components would usually be small. In other words, because strong aspects of meaning occur most often and are most likely to be right for their context when they do, and weak aspects do the converse, the average effect of "ambiguity" is small. Only when two aspects of a word's meaning are nearly equally strong (occur in equally important roles over equal numbers of paragraphs with very different meanings) will great conflict arise. Second, to the extent that an aspect of a word's meaning is unrelated to the rest of a paragraph in which it occurs, it is orthogonal to the paragraph's meaning to the same degree, and therefore acts only as noise.

To test this hypothesis, I studied LSA representations of multiple-sense words as defined by WordNet (Fellbaum, 1998; Landauer, 2002b). Every strongly multiple-sense word that we examined had significant similarity (cosine) to the text of each of its senses as defined in WordNet (with all forms of the word deleted from the definition.) Here is an example:

> *"Swallow"*—"The process of taking food into the body through the mouth by eating." cosine = .57
> *"Swallow"*—"Small long winged songbird noted for swift graceful flight and the regularity of its migrations." cosine = .30

However, there are exceptions to this picture within sentences. Sometimes word meanings affect each other. One example is in predication. In "My surgeon is a butcher" and "My butcher is a surgeon," different aspects of the meaning of a word are selected by differences in word order. Another example occurs in some metaphorical expressions, such as "His wife is the staysail of his life." Such phenomena—cases where a word does not correctly distinguish between meanings or lends only part of its meaning to a passage according to LSA—appear to occur much more frequently in linguistics books than in ordinary text. Nevertheless, they surely need explanation. Kintsch (1998, 2000, 2001, chap. 5 in this volume) has shown how LSA can help to construct explanations of this phenomenon.

For "My butcher is a surgeon" versus "My surgeon is a butcher," a set of n nearest neighbors to "butcher" in the semantic space are chosen. Vectors for

words among them that are sufficiently similar to "surgeon" are added to that for "surgeon." As a result, the sentence meaning emphasizes the aspects of the meaning of "butcher" that are also contained in "surgeon," but not vice versa. Kintsch has incorporated this idea into his construction integration (CI) model, giving the process an iterative settling effect that improves its outcome. The algorithm is tested by comparing the original and modified sentences with words or expressions that carry the meaning that was not in the receiving word previously and should have been magnified in the resulting sentence vector, for example "precise." Unfortunately, to date, the choice of which is the predicate and which the object still depends on human judgments. Artificial intelligence parsers go some distance here, but neither are good enough to do the job nor free of human intervention.

(Note here that strong word order effects are almost entirely within sentences. When LSA is used to measure the similarity of multisentence passages, word order effects become of less and less consequence because of the rapidly increasing dominance of word choice, as described later.)

Set phrases or idioms pose a different problem with a more obvious solution. Many such expressions may be thought of as patterns of words on their evolutionary way to condensing to a single word. Such patterns can be detected by the fact that they occur much more often than they would if the words were independent. These are commonly called collocations, and their identification has been done by several different statistical approaches. Once identified, collocations can be collapsed into a single word vector for participation in LSA space training. Adding such an algorithm is *ex cathedra* for LSA, but retains its spirit by eschewing direct aid from human knowledge of word meanings.

EVALUATIONS AND PROOFS

The initial demonstrations (Landauer & Dumais, 1997) of LSA's ability to simulate human word meaning made use of a standardized vocabulary test, Educational Testing Service's TOEFL (Test of English as a Second Language). The test presents a target word and four alternative words and asks the student to choose the one whose meaning is most similar. LSA was trained on a corpus of size and content approximating that of an average American college freshman's lifetime reading, based on a sampling by Touchstone Applied Science Associates (TASA) of books used in K–12 schools and found in their libraries. LSA took the same test and got as many right as successful applicants to U.S. colleges from non-English-speaking countries. Further simulations showed that the rate at which LSA acquired vocabulary as a function of the amount of language exposure closely approximated the rate of vocabulary growth of American children, approximately 10 words a day as measured by average gains over a year. Moreover,

just as is true for student learners, only 2 or 3 of the 10 words newly correct each day had been encountered during the last 24 hours. In LSA, the improvement came instead from the entailment of every word's relation to every other. Word meaning knowledge is not all or none, but grows gradually, not showing itself outwardly until good enough and in the right context with the right measuring instrument.

Recently, there have been several reports of models of different kinds that excel LSA on the same set of TOEFL items, getting as many as 90% correct, far better than the average student as well. These have all depended on having a much larger potential corpus to learn from, for example, the entire Internet, and searching it anew for the answer to each question. These models are of interest for practical applications such as data mining, and as different search techniques by which human memory might conceivably be used in performing the task. They also show convincingly that word meaning is latent in the evidence of experience and can be extracted from natural linguistic data. However, they do not speak to the question of whether such models can explain the human ability because they use more data than a typical human could have, and, especially, because they search for answers in a database after being given the question rather than answering from previously acquired knowledge. The fair comparison here, college applicants using the Internet to answer TOEFL items, would not be theoretically interesting for present purposes. Moreover, these models do not explain how words combine to form meaningful utterances, therefore nothing about meaning as construed here.

SYNTAX AGAIN

Some authors have also characterized LSA as a "bag-of-words" technique. This is true in the narrow sense that the data it uses does not include word order within passages. However, what the words are and what the model does with the words is critically different from the keyword or "vector space models" of current search engines with which the sobriquet of "bag-of-words method" is usually associated. In these techniques, query-to-document similarities are based on counting and weighting pair-wise word identities. The measurement of similarity of passages suggests throwing scrabble chips bearing words into bags of two different colors and counting (in some sophisticated way, of course) how many blue chips bear the same words as red chips. The result is just a value of the match in literal comparisons, no representation is first formed of the meaning of the things being compared, and there is no constraint on what words can be in the same bag. This is not a very appealing analog of the human process, much more like the naïve strawman notion of machine models of language attacked by Searle (1982) in his famous Chinese room allegory. By contrast, LSA accounts for the effects of all the words in each of the docu-

ments, matching or not, and such that their overall similarities match human judgments of semantic similarities. The LSA constraint that the combination of words in a "bag" add up to a meaningful passage for all passages in a very large sample of language gives a strong constraint on the content of a "bag." Thus, my fear that the use of "bag-of-words" for LSA significantly distorts its understanding for unwary readers.

BABEL AND THE EQUIVALENCE OF LANGUAGES

The severe dimension reduction from the original representation by, say, 100,000 words to 300 factors is an inductive step that forces the resulting vectors to drop unimportant detail to capture optimum levels of similarity. This process can be applied as easily to any language in which there are meaningful wholes composed of discrete meaningful units, be they French words or Chinese ideographs. This property is what has made it possible to automatically build LSA search engines with almost equal ease in Hindi, Arabic, Japanese, and Hebrew, as in those with Roman orthography. Non-English systems have also included German, English, Spanish, Italian, Swahili, and Latvian. It also makes it possible to build search engines in which passages composed in very different forms of meaning conveyance, such as Roman letters and Chinese ideographs, can be evaluated for similarity of meaning essentially as easily as passages all in one language.

All languages so far tried can be aligned with any other so that paragraphs in one have the same meaning as the other to nearly the same degree as those translated by expert human translators. Among other interesting things, this seems to confirm the belief that all languages are in some fundamental way the same. Because all languages must represent substantially the same set of meaning relations, they must also represent substantially the same relations among words. Importantly, however, in LSA the sameness is not given by observation of the phenomenon but by a well-specified computation that creates the equivalence.

The secret is again the common compositional constraint; in all languages, the parts add up to paragraph meanings. If two passages mean the same thing, then their summed vectors will be the same. Note the central importance of paragraphs (used in the general sense of a set of words with unitary meaning) as the determining unit of meaning. Informal support for paragraph meaning as the objective function of the learning model comes from observation of the normal units of discourse in which people write and converse, as appears in research corpora, and is taught as the optimal unit for a complete idea in composition classes.

Note now that the solution of the system of linear equations is unique only up to linear transformation and there are therefore an infinite number of solutions. Thus, even if two individuals had almost identical experience,

small variations and differences in their LSA-like brains might result in quite different solutions, as might different languages.

How then can people understand each other? The theoretical answer is that every solution would approximate a linear transform of every other, and a mechanism exists by which we align the transformation that takes an individual's semantic space to some particular language's statistically canonical cultural form. The LSA answer, of course, is that closely matching a small fraction of my words and utterances with yours will drag the rest of the structure with it, as in aligning two maps by overlaying just two points (see Goldstone & Rogosky, 2002, for a related computational technique).

An LSA-based developmental hypothesis might go like this. Babies first learn a primitive embedding structure of word–word relations by hearing words in multiple verbal contexts, then gradually add mutually consistent words and word groups to an evolving mini structure of meaningful interrelations. Such growth will resemble that of a crystal; immersed in the medium of words and passages, new words will attach where meaningful combinations are ready to use them, and new ability to understand word combinations will emerge as more words take their places.

Word and passage meanings should start out quite ill or fuzzily defined because of the sparseness of possible embeddings. Early vocabulary should consist primarily of words of the highest frequency in the child's attentionally filtered verbal experiences. Most of this would come from the ambient speech to which they are exposed (Hart & Risley, 1995).

Ambient language exposure consists primarily of utterances by competent speakers in phrases, sentences, and paragraphs, which the child can begin to understand slowly and again fuzzily. By LSA, the quality of the representation of words and paragraphs will increase in a mutually reinforcing iterative process. As simulated by LSA, the lion's share of this incremental growth will be hidden from easy detection. Learning to construct meanings of words and passages are tightly coupled and follow a similar course. Eventually, there will be a sufficient core to support vocabulary growth at the 10 per day observed to pass the TOEFL threshold at age 12. Note that the points at which this process includes perceptual and motor context (again for both positive and nonoccurrence effects) are continuous, simultaneous, and intermingled with language experience from the beginning. The meaning of "tree" becomes better defined as a better place for it is constructed in the semantic space, and a better place is constructed as the learner has more experiences containing and not containing a perceptual tree.

In an unpublished pilot study, Dumais and I examined the trajectory of LSA word neighborhood changes during early learning. This was done by substituting nonsense words for actual words with increasing numbers of occurrences. At first, a word had many word neighbors of modest similar-

ity, mimicking the early errors of overgeneralization that children make. Then there was a contraction to a small number of more tightly clustered neighbors, and finally to an again larger number, but with both closer neighbors and presumably neighbors that matched the multiple senses of the words—although the last property was not investigated at the time.

In Landauer and Dumais (1997), a hypothetical course of language experience is presented for a child to learn what "hurts" means when uttered by its mother. The new knowledge arises from the word's embedding in other utterances that put it near the child's own learned representation of "hurts," which by the cultural conversion presented previously, makes it the same as the mother's. This would also be the explanation of Quine's famous objection to association as the mechanism of word learning (Quine, 1960; see also Bloom, 2000). A perceived object fits into a visual/semantic space by its LSA similarities to things already there (see Edelman, 1998, 1999; Valentine, Abdi, & Otoole, 1994).

Much the same mechanism is taken advantage of in the (proprietary and successful) cross-language information retrieval method described earlier. The system creates separate semantic spaces that share a relatively small number of documents that are known to have close to the same meaning, for example, translations or news accounts of the same event from different language sources. It places those known to be alike in the same relation to one another in a new joint semantic space, then moves the rest in the same way. This never places a document or word type from one language exactly in the same place as any from another (in English and Chinese a "component" usually has no very similar vector in the other language). Most words considered to correspond in bilingual dictionaries will tend to be quite close to each other, and documents that have been carefully translated will be very close to each other.

This also yields an explanation of why the best way to learn a new language is by immersion; a small number of common words, a small amount of direct instruction and experience and/or of hearing foreign words or passages in situations where it is obvious how the meaning would be expressed in L1 will support alignment of L1 and L2. It also mirrors the common intuition that different languages do not map perfectly onto one another, and there may often be no way to express the same idea exactly in a different language.

Moreover, if the relations among percepts—both innate and experiential—were organized in the same manner, it would take only a comparatively few correlations of perceptual and linguistic experience to make all their connections fall in line to a close approximation. A Helen Keller could put unseen and unheard objects and language together on the sole basis of correlations between touch, smell, and taste stimuli along with a very brief early history of normal perceptual-language associative experience (Keller, 1905).

Because everybody shares experiences with many other people, statistics will insure that there is good, if imperfect, agreement across members of a community. This would happen by a process of consensus promoted by both interpersonally engaged and ambient conversation (and, recently, newspapers, popular songs and books, movies, and television), so that semantic knowledge and abstract ideas will recursively feed on themselves. Of course, each person's understanding of the meaning of a word will still be slightly different, in ways ranging from minute to large, for example, if one person has read a word only in one "sense" and someone else has read it only in another.

Such a process remains to be simulated in detail to see if it would have the power and reach needed, but at least conceptually it provides a possible solution to a chronic philosophical problem referred to as publicity, how people share meanings as they must.

One piece of supporting simulation data was reported by Landauer and Dumais (1997). The rate of acquisition of new word knowledge was simulated as an accelerating function of the number of words previously encountered. Whereas a simulated 50-year-old reader would learn a new word in two encounters, it would take a 20-year-old person eight chances. Of course, a person's total vocabulary follows a typical S-shaped growth curve, the rate of growth first increases, then slows down as the number of unknown types encountered decreases.

The LSA prediction is that vocabulary should grow in each individual by embedding words both old and new in large, common, and increasingly stable semantic space, allowing people of all ages to continue to improve their sharing of meanings with others. The principal evidence supporting this expectation is that addition of new paragraphs to an LSA information retrieval system requires less and less frequent re-computation of the space to give words and passages appropriate meanings. After learning from a large corpus, newly encountered words can be "folded in," that is, placed at the average point of all the paragraphs in which they occur without re-computing the SVD, thus obeying the fundamental LSA constraint. Unless there has been a relatively large addition, say greater than 20%, of new words or a significant change in the corpus domain, the difference between this way of adding vocabulary and that of recomputing the SVD is negligible.

FINAL WORDS ABOUT WORD ORDER AND MEANING

LSA's successes would seem utterly impossible if word order played as dominant a role in the actual use of verbal meaning as it does in the science of linguistics. How can this be explained? Here are three approaches: See if we can estimate just how much is missing by ignoring word order, try to put bounds on the relative contributions of word combinations and word

order to passage meaning, consider what other roles the omnipresence of word order conventions in many languages might play.

It is worth noting that many, perhaps most, languages are not nearly as fussy about word order as English, and their informal speakers are not nearly as fussy as their teachers and theorists. One reason is that some of the information carried by syntax in English is carried in some other languages by a greater variety of differential word forms and affixes that index the sentential roles of words rather than order dependant combinations.

An informal example may help intuition. Readers will have little trouble figuring out what the following word string means:

["order syntax? much. ignoring word Missed by is how"]

Scrambled sentences and passages are often fairly easy to understand, even without laborious reconstruction. By first getting the topical gist from the combination of words, then rearranging most of the words to grammatically correct and empirically probable orders, one can usually recover the original form of an utterance, at least modulo differences that do not affect meaning. It is not always trivial to construct a paragraph that when randomized defeats general gist understanding or leaves serious ambiguities in the mind of the reader. Even double, missing, and incorrectly placed negations and modifiers, as well as outright sentential contradictions in paragraphs, frequently go unnoticed by readers (Kintsch, 1998), and their lack often does not appreciably alter the meaning of the text for the reader. The LSA interpretation is that the meaning of a passage being the average of many words, the effect of a few deviant meanings may sometimes have little effect on the whole.

We can go beyond this qualitative argument to a quantitative estimate in the following way. We used LSA alone in a linear ordering algorithm to place a large set of essays on a line such that the 300-dimensionality similarities between them were minimally distorted. We then compared this ordering to one based on average scores for the same essays given independently by two expert humans. Finally, we measured the amount of shared information in the scoring of the essays (a) by the two human experts—presumably based on all they could extract from all aspects of the writing, including syntax—and (b) between the scoring by LSA and the humans. The measure used was mutual information (also known as cross-entropy). This measures the amount, in information-theoretic bits, by which the uncertainty (entropy) in one source can be reduced by knowing the other. The result was that the human–human mutual information was .90 and the average machine–human was .81. That is, the machine, on average, shared 90% as much information with each of two human experts as the experts shared with each other. This gives us a first rough esti-

mate that 10% of information in multisentence texts that is used by humans comes from word order. It would be hazardous to make too much of this result without replication and confirming extensions, but it is evident that at least in judging essay content quality, the opportunity to use word order does not greatly improve expert performance.

The next approach is more abstract. For convenience, assume that a typical well-educated adult English speaker knows 100,000 words well enough to understand their contributions to the meaning of sentences (see Landauer & Dumais, 1997), and an average sentence contains 20 words. The number of possible combinations of words in a sentence is then $100{,}000^{20}$, the number of information-theoretic bits $\log_2 (100{,}000)^{20} = 332$ bits. The number of possible orders of 20 words is 20!, the number of bits $\log_2 (20!) = 61$ bits. The total maximum information in a 20-word sentence is thus 332 + 61 = 393, of which 61/393 = 15.4% is from word order. If we add in the corresponding amounts for a series of sentences in a paragraph or essay, then the situation gets even more lopsided because possible word combinations multiply across multiple sentences and paragraphs, whereas the number of permutations only add, word order effects being almost exclusively within sentences. Thus, this approach comes interestingly close to the first one, with 10%–15% of information in English text from word order.

In many of the practical applications of LSA, people have joined it with statistical models of word order. These methods, which are the workhorses of modern speech recognition systems (Rosenfeld, 1996), model language by computing the frequency with which sequences of word—*n*-grams—usually 2–5 in length, appear in large corpora of representative text. Such models can be somewhat more powerful when applied to text than to speech because in text what follows can affect the comprehension of what came before, whereas in real-time speech processing the previous words are too soon gone. Nonetheless, about all they have been made to do is to tell us is how likely it is that the observed order of words is expectable in general or in a certain domain or from a certain source. Nonetheless, they can provide yet another hint about the limits of the effect of word order. In an unpublished 1996 pilot study, Noah Coccaro and I selected random 10-word sentences from the *Wall Street Journal*. The order of words in each sentence was then randomly scrambled. Finally, we tried to recover the original order by using *n*-gram probabilities from a large *Wall Street Journal* corpus to find the word order that had the highest probability. About half of the sentences were perfectly recovered, another quarter sufficiently that no change in meaning resulted, the rest with minor ambiguities.

Finally, let us speculate a bit on why English, and to a lesser extent other language speakers, bother themselves, their listeners, students, and editors so closely and insistently about adhering to conventional patterns. Please note that I am not asserting that word order is not important to meaning; it

clearly is. What I wish to point out, however, is that its role in verbal meaning may have been overestimated, and thus the importance of word combination underappreciated.

Clearly, one source of the ubiquity of word order conventions is a matter of style, like wearing skirts or ties or serving the dessert before or after the cheese. Another is plain cultural habit. One of your ancestors one day started usually putting adjectives before nouns, and her children learned *n*-gram probabilities from her. Read some literature or science from the mid-19th century (Darwin is nice for the purpose), and you often find the word order consistent but consistently different from Pinker's. To what extent did formal English evolve a more efficient way to represent and convey meaning through word order, and to what extent was there just the same sort of evolutionary cultural drift as there was in hats, driven in the same way as Peacock's tails?

I think fashion and habit are part of the story, but there is yet another reason to use conventional word order even supposing it had no influence on meaning. The reason is increased efficiency and accuracy for speaker, hearer, reader, and thinker. Such an efficiency would accrue to almost any conventional word order. To illustrate this, imagine a language that was totally without word order conventions. To say anything you would just choose a set of words and string them together in any order. This language would require a rather huge number of words, but many real languages, such as German, are more forgiving about word order than English, and some encryption schemes employ orderless code blocks. This is accomplished by using more complex words—making the building blocks carry more information—and partly by more flexible conventions. Given a partially order-free language—for example, only within sentences—you would soon find that if you put your words in alphabetic order when speaking or writing your hearers or readers would make fewer mistakes in understanding; you would thereby have created an error-correcting coding scheme. Note that you would also think, speak, and understand faster and more accurately because each word would cue the next.

The proposal, then, is that conventional (rule governed maybe, but I prefer experiential and statistically shaped) word order provides a communicative advantage. Extending the speculation, the universal grammar or UR syntax that Chomsky and followers have sought is a natural consequence of agreeing on what kinds of meanings should come after which others to make talking more effective, and various interacting groups have stumbled on and evolved various ways to do it. Unfortunately, linguistic theory has yet to find a computational theory that can simulate that evolutionary process or the passing of its result from old to young through exposure to language.

The critical problem of utterance production—speaking or writing—raises its intransigent head here. LSA theory and systems are not de-

signed to simulate or explain this essential property of language. The theory says that passages that are emitted should obey the constraint, but this is of limited help. What mechanism could manage to emit words, sentences, and paragraphs that do that? Some unknown (to LSA, but partially described by phenomenon-level linguistic theories and rules) computation must insure that utterances usually create comprehensible word strings.

In addition, the meanings expressed in sequential mathematical and logical statements are not in LSA's theoretical purview, but models of them can nevertheless profit from using LSA as a component. For example, "John hit Mary" might be decomposed in some propositional form in which much of the meaning is order free, for example [a hitting < {[John, Marry]} < (hitter: John)]. However, there is still order in deciding who is the hitter. Kintsch's work with predication, metaphor, and analogy models (chap. 5 in this volume) takes this tack, marrying LSA to automatically represent individual word meanings with syntactic models that are effective but rely on help from human coding.

As with every other scientific theory, LSA succeeds by abstracting only a limited range of phenomena to explain from the enormous complexity of nature. LSA offers one explicit and computable mechanism for an essential and previously inexplicable component of language, how the meanings of words and passages can be acquired from experience. Perhaps LSA's success for this component will encourage new research into how other important properties of language that must arise from normal experience—the syntax of one's language in particular—do so.

SOME FINAL PHILOSOPHICAL AND PSYCHOLOGICAL MUSINGS

The property of LSA that a passage of words on being understood, turns into a something that cannot be turned back into its words is familiar to both intuition and experimental psychology. Kintsch calls this the transition from text base to situation model. The abundant psychological evidence is that people are hard-pressed to recall verbatim what they have read a few occupied minutes earlier, but can recognize it as familiar when seen again. More wonderfully, they retain information that the text conveyed and can recognize it in hundreds of disguises and use it in hundreds of ways: to paraphrase, to reason from, to piece together with others of the same, to criticize and praise. More recent research has shown that knowledge of the original words is not completely lost; give a subject a choice of those that were there or not and they will do quite well. LSA at least shadows this. Even though the exact words are irretrievably mingled into something else, a word originally there is much like the new whole. We have built algorithms that can generate small sets of words, either the originals

or others, that carry the core of the original meaning, such that added together they produce a vector nearly identical to the original.

However, the fact that passage-meaning vectors, and even whole essay vectors—which are essentially sums of passage meanings—cannot be turned back into words has interesting implications. We may think of a passage vector as an unexpressed idea. Consider the everyday introspective experience that a meaning is there before it becomes words. This is demonstrated by our claims that we cannot find words to express an idea, that we did not mean what we said, and by our ability to edit a passage to make it express more nearly what we meant.

Words themselves are discreet isolates. What are they when combined and forgotten? Have they lost their individuality? LSA suggests an answer. First, let us observe that people rarely speak isolated words, they usually speak and understand groups of words, and typically groups of words larger than a sentence, more akin to paragraphs (this is another reason that constructing paragraph meaning is so important in LSA). A number of interesting conjectures spring from this observation. First, we may suppose that the comprehension of a new sentence or passage consists not only of the addition of word meanings but also of similar nonverbal vectors, including perceptual records, retrieved from memory. Thus, reading comprehension would be enriched by the recovery and incorporation of associated information.

Perhaps unconscious thoughts of all kinds are just "untranslatable" vectors derived from perceptual experience and discourse, additive combinations of different passages that cannot be unpacked into strings of words. Might not such "nonverbal" thoughts nevertheless constitute an important part of cognition? Surely the content of articles and books read long ago still influences thought, decision, and action even though the individual words are not recoverable. Perhaps the residues of condensed unverbalizable, but nevertheless influential, memories play a very large role in both our conscious and unconscious life, biasing verbally expressible thoughts, feeding unverbalizable ones and their expression in emotion and action. Maybe falling in love relies on a match of untranslatable vectors, which is why a matching verbal checklist does not always equal romantic chemistry. Kintsch (1974) made great progress in this same direction by proposing that word strings were turned into a form of logical propositions for storing in memory, a strategy often adopted in artificial intelligence. In these approaches, however, literal words or equivalent discrete symbols arranged in structured formats still carry the meaning. What is proposed here is not that such representations do not exist, but that much of cognition may consist of LSA-like representations that carry and combine meanings in a very different way.

Other phenomena of mind seem to find possible explanations in this manner. Where do thoughts and new ideas come from, from additive com-

binations of other thoughts all verbally inexpressible or only partly so? From the averaged vectors of two or more words that do not correspond to any single word? How about the unverbalizeable emotional storms of depression and anxiety, the word-salads of schizophrenics? Are meaning vectors making trouble without being able to speak? How about the mysterious origins of insights, intuitions, sudden solutions to math problems? Are meaning vectors doing wonders without telling? Think about consciousness. Maybe some of what we think of as not conscious is just stuff we cannot talk to ourselves about even though it may be quite full of complex meanings. Maybe the "meanings" of sunsets, thunderstorms, gestures, or of supernatural beliefs, irrational fears, Freud's unconscious wishes, automatic actions, unintentional learning, the search for the meaning of life, can all find new explanations in the dynamics of verbally inexpressible LSA-like vectors.

IN SUM

LSA demonstrates a computational method by which a major component of language learning and use can be achieved. It is in that sense that it is a theory. It is specifically a theory of meaning because it applies to and offers an explanation of phenomena that are ordinarily considered to be manifestations of meaning: the expression, comprehension, and communication of ideas and facts in words and passages of words. It makes possible computer systems that accomplish a wide range of cognitive tasks performed by humans, and often does them essentially as well and with the same detailed performance characteristics. This makes its basic mechanism, or something equivalent to it, a candidate for explaining the corresponding human abilities. The research strategy and program that the LSA community follows is well described by Stokes in *Pasteur's Quadrant* (1997). Start with a practical problem, do the science needed to understand what's going on and how to fix it, test your understanding and its completeness by whether you succeed and how you fail: iterate.

LSA's initiating event was people's difficulties in finding services they wanted in the Bell System Yellow Pages. Observational experimentation discovered that the cause was that there were always many more words with related meanings that searchers tried than indexers indexed (Furnas, Landauer, Gomez, & Dumais, 1987). The solution was to find a way for a computational model to learn word meanings from vast amounts of exposure to text, just as humans do, so that it could tell when an inquiring person's words meant nearly enough the same thing as its. The tests were manifold, some abstracted controlled laboratory experiments, many more by building software systems that had to understand the degree to which two words or passages had the same meaning. The model did surprisingly well, underwriting many useful

inventions, insights into the nature of verbal meaning (word choice more important relative to word order than previously suspected), new theoretical conceptions of how language might work (passage meaning as a sum of word meanings), and realizations of where the theory is incomplete or falls short (accounting for the effects of word order, analogy, inference, generation of meaningful language). What LSA can do has pointed to a new path in the study and engineering of meaning.

ACKNOWLEDGMENTS

Research was supported by the NSF, AFRL, ARI, ONR, DARPA, IES, McDonnell Foundations. Thanks to Touchstone Applied Science Associates.

REFERENCES

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bartlett, F. C. (1932). *Remembering*. Cambridge, England: Cambridge University Press.

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Applications*, *6*(1), 13–49.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM: Review*, *37*(4), 573–595.

Bickerton, D. (1995). *Language and human behavior*. Seattle, WA: University of Washington Press.

Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT Press.

Chomsky, N. (1991a). Linguistics and adjacent fields: A personal view. In A. Kasher (Ed.), *The Chomskyan turn* (pp. 3–25). Oxford, England: Blackwell.

Chomsky, N. (1991b). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed.), *The Chomskyan turn* (pp. 26–53). Oxford, England: Basil Blackwell.

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.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*(2), 145–193.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, *23*(2), 229–236.

Dumais, S. T., Landauer, T. K., & Littman, M. L. (1996). Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In SIGIR'96 - Workshop on Cross-Lingusitic Information Retrieval, (pp. 16–23).

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences*, *21*, 449–498.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2005). *Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia.* Manuscript submitted for publication.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *97*(2), 11885–11892.

Fellbaum, C. (1998). Introduction. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 1–19). Cambridge, MA: MIT Press.

Fodor, J. A. (1987). *Psychosemantics*. Cambridge, MA: MIT/Bradford.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Analysis of text coherence using latent semantic analysis. *Discourse Processes*, *25*, 285–307.

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

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*, 379–401.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447–474.

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*, 295–320.

Griffiths, T., & Steyvers, M. (2003). Topic dynamics in knowledge domains. *Proceedings of the National Academy of Sciences, 101(*Suppl. 1), 5228–5235.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes.

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

Keller, H. (1905). *The story of my life*. New York: Doubleday, Page & Company.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, *7*, 257–266.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Laham, D. (1997). Latent semantic analysis approaches to categorization. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (p. 979). Mawhwah, NJ: Lawrence Erlbaum Associates.

Laham, D. (2000). *Automated content assessment of text using latent semantic analysis to simulate human cognition*. Unpublished doctoral dissertation, University of Colorado, Boulder.

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Landauer, T. K. (2002a). On the computational basis of cognition: Arguments from LSA. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 43–84). New York: Academic Press.

Landauer, T. K. (2002b). Single representations of multiple meanings in latent semantic analysis. In D. Gorfein (Ed.), *On the consequences of meaning selection* (pp. 217–232). Washington, DC: American Psychological Association.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automated essay assessment. *Assessment in Education: Principles, Policy and Practice*, *10*(3), 295–308.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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 sciences society* (pp. 660–665). Pittsburgh, PA: Lawrence Erlbaum Associates.

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Osherson, D., Stob, M., & Weinstein, S. (1984). Learning theory and natural language. *Cognition*, *17*, 1–28.

Pinker, S. (1994). *The language instinct*. New York: HarperCollins.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive sciences. *Behavioral and Brain Sciences*, *3*(1), 111–169.

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

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.

Ridley, M. (1994). *The Red Queen: Sex and the evolution of human behavior.* London: Penguin.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, *10*, 187–228.

Searle, J. R. (1982). The Chinese room revisited. *Behavioral and Brain Sciences*, *5*, 345–348.

Skinner, B. F. (1957). *Verbal behavior*. East Norwalk, CT: Appleton-Century-Crofts.

Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Valentine, D., Abdi, H., & Otoole, A. (1994). Categorization and identification of human face images by neural networks: A review of linear autoassociative and principal component approaches. *Journal of Biological Systems*, *2*, 412–423.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, England: Blackwell.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., et al. (1998). Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, *25*, 309–336.