On the computational basis of learning and cognition: Arguments from LSA.

Thomas K Landauer

To deal with a continuously changing environment, living things have three choices. (1) Evolve unvarying processes that usually succeed. (2) Evolve genetically fixed (possibly ontologically developing) effector, perceptual, and computational functions that are contingent on the environment. (3) Learn adaptive functions during their lifetimes. The theme of this chapter is the relation between (2) and (3): the nature of evolutionarily determined computational processes that support learning. Examples of this focus are neural mechanism conjectures, connectionist modeling, and mathematical learnability theory. The somewhat different approach taken here is to ask what evidence about the world we have access to and what can be done with it. This chapter cannot offer an exhaustive and rigorous treatment of the issue. It presumes only to present an example of how its consideration may lead to interesting results and insights. Its main point is to argue that learning from empirical association, if done right and writ very large, is capable of much more than often supposed.

What kind of evidence do we animals have from which to induce knowledge? Mostly observational. We use our limited effectors and perceptors to explore and learn how objects and events are related to each other. There is some opportunity to arrange what will be observed by being in the right places at the right times. And there is some opportunity to experiment; to try to affect the world and see what happens. However, the principal evidence we get is patterns of stimulation as they occur over time. The vast preponderance of our available raw data is empirical association: the occurrence of two or more perceptual or effective elements in time: that is, coincidence, co-occurrence, contingency, or correlation.

The question is for what can this kind of data be used? The British empiricists had it right in pointing out that associational data must be the fundamental basis of acquired knowledge, because there isn’t anything else. But they did not (probably could not, given the theoretical tools

of the day) rigorously work out what could and couldn't be done with it. In fact, Locke's postulation of similarity as a separate kind of association data, in addition to contiguity, implicitly assumed, but left unexplained, a computational mechanism by which perceptual data are combined and compared. A critical missing piece was the basis of similarity, especially similarity that must be learned, for example that between a car and a truck, or the words "man" and "wife". As Goodman (1972) put it, similarity taken as primitive is an imposter, a fact in need of explanation, not an explanation. How we compute similarity is one of the essential questions in understanding how the data of experience are made useful. It is intimately related to object identification and recognition, whether innate or learned, and to generalization, categorization, induction, prediction, and inference.

In this chapter, I take two psychological phenomena as cases for illustration and discussion, visual object recognition and verbal semantics. I find the questions and potential answers in the two cases remarkably similar, and the counterpoint of their joint discussion illuminating. The fundamental questions in both cases are what the elements of association are and how they are combined into useful representations and thoughts. However, exploring the problem by computational simulation has been easier and more revealing for verbal meaning, so my strategy is to describe some of what has been learned in that way, and then to consider how the lessons might apply to object recognition.

The discussion is organized in a somewhat spiral fashion. I first raise the issue of the nature of the basic elements of empirical association and illustrate it by the case of learned object recognition. This leads to the hypothesis that the choice of optimal elements may provide a relatively small part of the solution of the problem; what is done with the co-occurring elements appears to be more important. I then move to the learning of word and passage meaning because this domain exhibits the problem in a manner that is particularly convenient to model; we can give a computer the very same mass of perceptual input that literate humans use for much of their learning. I first show how a different kind of co-occurrence data and a different form of

computation can yield much more knowledge than has usually been supposed (e.g. by Bloom, 2000, Chomsky, 1991 a & b, Gleitman, 1990, Osherson, Stob and Weinstein, 1984, Pinker, 1994). The co-occurrence data is not of words and words, but of words and contexts. The computation is not estimation of the probability of sequential contingencies between words, but rather the use of the algebra of simultaneous equations to induce meaning relations between words from all the contexts--not only those shared--in which they do and do not appear. Next, I explain how these ideas are implemented in the Latent Semantic Analysis (LSA) learning model through singular value decomposition, an efficient matrix algebraic technique that represents words and passages as high-dimensional vectors in much the way that neural nets represent inputs and outputs by values on hidden units. I then list a variety of human verbal comprehension performances that LSA simulates surprisingly well: for example, that it passes multiple choice final exams in psychology and equals experts ability to score the conceptual content of student essays on substantive topics.

Importantly, LSA's success depends critically on sufficient scale and sampling, on the amount and naturalness of the data that it is given. Its ability to represent word meaning comes from relating all of tens of thousands words and tens of thousands of contexts to each other in a mutually consistent manner. It's success also depends on choosing the right number of underlying dimensions of meaning to extract. Representing the similarity of millions of local observations by values on many fewer dimensions induces an enormous amount of "latent" information. This makes it possible, for example, to compute that two passages have the same meaning even if they contain no words in common.

The lesson I then take from LSA's successes is that empirical association data, when sufficient to accurately induce how all of its elements are related to each other, makes learning from experience powerful enough to accomplish much, if not all, of what it does. The next main section of the chapter conjectures about how the same principles might apply to object

recognition. The aim is not to propose a worked out model, but to suggest that the scope of the general principles may be quite wide indeed.

The chapter concludes by discussing the implication that relations between thoughts, words, passages, objects, events, and intentions may all be induced by the same fundamental process. Along the way, some words are also spent on the varieties and limitations of current instantiations of models of the kind proposed, some objections to LSA that have been raised by others, and the future needs and prospects of this line of research and theory.

THE ELEMENTS OF ASSOCIATION.

If we are to learn from empirical association, we need things to associate. To start off with, some primitive elements given by nature. What are these?. Are they discrete, atom-like elements, or values of continuous variables? If discrete, are they ordinal valued, and if so, how many different values do they have? Do they come with structure, relations to each other, ready-made similarity in the sense that they can substitute for one another in mental computation? How many different variables are there? Are they independent or correlated? How freely do they combine; do they have constraints or favoritisms like molecules? Orderly attachments like crystals or proteins? Other innate structure? How much contextual invariance do they display? A closer look at these questions is provided by considering how object recognition might be performed by animals or machines.

Object recognition

Perhaps the most central and difficult question about object recognition is how a three dimensional object can be recognized as the same, and discriminated from and related to others, despite changes in viewpoint, lighting, and context. Actually there is a prior question; how can an object be recognized and distinguished at all? The simplest theory, perhaps, is the template. Make a spatial record of the object’s shape as it appears on the retina or in the visual cortex. Move the record over incoming scenes to try to match it point for point. Measure the degree of match. The

orientation and context problems are attacked by trying all positions, parts, and magnifications of the template, the lighting problem by normalizing with respect to overall illumination. Pursued with great diligence and cleverness, machine implementations of this approach work fairly well for some purposes, such as recognizing printed addresses viewed at the perpendicular. For irregular solid objects seen a few times, then rotated in three dimensions, it fails badly. But this is a feat that most animals can do quite well.

What elements of experience are assumed in the template hypothesis? Are they tiny points, corresponding, say, to retinal receptors? The smaller the points, the more accurately they will have to be aligned for any old sets of points to match any new ones, but the fewer false positives they will necessarily generate. By making the points bigger and fuzzier we can get more generalization; small translations or rotations may leave the match better than competing patterns, but at a cost to precision.

This is an example of the first possible solution to the unit issue in using co-occurrence data; use a unit that is an optimal compromise between precision and forgiveness. This solution sometimes goes by the names of grain size or coarseness of coding. Unfortunately, while helpful, it is clearly insufficient. A rotated irregular 3-D object will not often have a pattern, no matter what size and fuzziness the spots, that overlaps discriminatingly with the original.

The next possibility is to replace fuzzy spots with a set of shape-diagnostic and view-, rotation-, illumination-invariant features. A well-known set is the few dozen “geons” proposed by Biederman (1987). These constitute an alphabet of proposed elementary shapes that are claimed in combination to be sufficient to characterize any object in any view so as to differentiate it from any other with accuracy approaching that of human perception. Edelman (1999) reviews evidence and arguments that geons are not a complete solution, for example that no one has come close to automating object recognition with human-like generality by this approach, primarily because the scheme does not yield adequate interpolation and extrapolation to new viewpoints or occlusions. Nevertheless, something of the kind is a necessary component of the solution. We need elements

to combine, and the more discriminating and invariant they are the better. However, it is apparent that good units alone are not the solution, that we will need more, and that the more will lie in the combining computations.[1]

Language

Where visual object recognition can be conceived of as a sequence of operations on stationary input patterns--one camera image or saccadic scene at a time--the fundamental elements of spoken language are ephemeral patterns defined by continuous temporal variation. Printed language discretizes the acoustic stream and turns it into a sequence of visual objects. A hierarchical organizational scheme apparently characterizes all languages. The acoustic stream is partitioned into easily articulated and recognized units (phonemes), sequential strings of these into clusters of various sizes (consonant-vowel clusters, syllables), and these into very large alphabets of discrete words, which are combined into a virtually unlimited number of objects (idioms, phrases, sentences, utterances, paragraphs). The language learner needs to be able to recognize and differentiate units at all these levels.

We will leave aside how people learn to recognize the physical word-form and sub-word-form units, only mentioning that from phoneme up they all must be mostly learned because they differ from language to language, whereas we can imagine that innate units are useful farther up the corresponding hierarchy for vision, essentially up to units useful in the shared environments of humans over the last ten millennia. The reason for omitting the lower levels here is pragmatic. Exploratory simulation of the mechanism for combining words into utterances can take advantage of immense samples of natural language already transcribed into word units and the ability of computers to deal with them. The form of that recoding is completely natural. It is not a hypothetical representation of input features created to serve the needs of a theory, but the actual input code for much of human use and learning of language. To repeat, the hope is that things discovered here will have relevance elsewhere.

COMPUTATIONS FOR COMBINING ELEMENTS

One strategy for developing theory about a natural process is to start with very simple assumptions or models and see how much they explain before introducing what may turn out to be unnecessary complexity. For combining words into meaningful utterances, perhaps the simplest model to consider is the unordered set of word tokens in a passage. In a printed passage each string of characters separated by a space or punctuation-mark may be taken as a word token. It is helpful to start this analysis by calculating the potential information in the combination (in the mathematical sense) of words and in their order (their permutation), respectively. To keep the numbers simple, assume that comprehension vocabulary is 100,000 words, that sentences are 20 words long, and that word order is important only within sentences. Then the contributions, in bits are $\log_2 (100{,}000)^{20}$ and 20! respectively, which works out to over 80% of the potential information in language being in the choice of words without regard to the order in which they appear. Using this observation to justify ignoring word order permits a convenient simplification. We assume that the elements are additive. As we will see, it turns out that this “bag of words” function, if properly realized, can produce surprisingly good approximations.

Learning word meanings and knowing passage meanings

Here is the way we go about it. The object that we want to account for in language is a transmittable or thinkable meaning. It is obvious that all the meaning of a passages is contained in its words, and that all its words contribute to its meaning. If we change even one word of a passage, its meaning may change. (The meaning of a passage plus its context is not, of course, contained in just the words, as Bransford and Johnson (1972) pointed out.) On the other hand, two passages containing quite different words may have nearly the same meaning. All of these properties are neatly represented by assuming that the meaning of a passage is the sum of the meanings of its words.

$$\text{meaning of word}_1 + \text{meaning of word}_2 + \ldots + \text{meaning of word}_n = \text{meaning of passage.}$$

Given this way of representing verbal meaning, how would a learner go about using data on how words are used in passages to infer how word meanings and their combinations are related to each other? Just assuming that words that often occur in the same passages have the same meaning won't do at all. For one thing, it is usually false; it is the combination of words of different meanings that makes a passage meaning different from a word meaning. Consider the following passages, which are represented as equations as specified above:

System 1.

ecks + wye + aye = foo

ecks + wye + bie = foo

Ecks and *wye* always co-occur in the same passages, *aye* and *bie* never. Together the two equations imply that *aye* and *bie* must have the same meaning, but nothing at all about the relation between *ecks* and *wye*. Thus, the way to use empirical association data to learn word meanings is clearly not just to assume that words that have similar meanings to the extent that they tend to appear together.

Now add two more equations.

System 2.

ecks + wye + aye = foo

ecks + wye + bie = foo

ecks + wye + cee = bar

ecks + wye + dee = bar

We now know that *cee* and *dee* are also synonyms. Finally consider:

System 3.

aye + cee = oof

bie + dee = rab.

To be consistent with the previous passages, from which *aye* = *bie* and *cee* = *dee*, these two passages must have the same meaning (*oof* = *rab*) even though they have no words in common.

Here we have the makings of a computation for using observed combinations of elements that appears more subtle and promising than simple classical conditioning in which one stimulus comes to act much like another if and only if it occurs soon before it, or by which two passages are similar just to the extent that they contain the same words (or even the same base forms).

The next step is to formalize and generalize this idea. Doing so is quite straightforward. Consider every passage of language that a learner observes to be an equation of this kind. Then a lifetime of language observation constitutes a very large system of simultaneous linear equations. This set of equations is certain to be highly "ill-conditioned" in mathematical terminology, meaning that there will be too few equations to specify the value of many of the variables and some of the subsets of equations will imply different values for the same variable. As a model of natural language semantics, these deficiencies do not seem out of place; word and passage meanings are often vague or multiple. Mathematically, such complexities can be dealt with by abandoning the requirement of finding absolute values, settling for relations among the variables, and representing them in a richer manner than as real values on a number line (scalars.) One computational method for accomplishing this is called Singular Value Decomposition (SVD)[2]. SVD is a matrix algebraic technique for reducing the equations in a linear system to sums of

multidimensional vectors. Good introductions to the mathematics may be found in Berry (1992) and Reyment and Jöreskog (1996) and its original use in language modeling in Deerwester et al. (1990).

The Latent Semantic Analysis model (LSA) uses SVD to simulate human learning of word and passage meaning. The first step is to assemble a corpus of natural language that is as similar as possible in size and content to that to which a simulated human would have been exposed. The corpus is parsed into meaningful passages such as paragraphs. A matrix is formed with passages as rows and words as columns. Each cell contains the number of times that a given word is used in a given passage. A preliminary transform is customarily applied to the cell values to change them into a measure of the information about passage identity that they carry, a transform that resembles first order classical conditioning of two stimuli (in this case a word and its passage context) as a function of occurrences in multiple contexts (Rescorla and Wagner, 1972). SVD is applied to re-represent the words and passages as vectors in a high dimensional "semantic space". The solution corresponds to the system of equations postulated in the model in that the vector standing for a passage is the vector sum of the vectors standing for the words it contains. In LSA, the similarity of any two words or any two passages is usually computed as the cosine between them in the semantic space; words or passages that are identical in meaning according to the model have cosines of 1, unrelated ones, 0, and ones of opposite meaning (which never occur in natural languages), -1.

New SVD algorithms for very sparse matrices (> 99.9% of the cells in an LSA matrix are typically empty) coupled with high performance computers with great amounts of memory can now compute SVDs for matrices of >100,000 words by a million passages in > 400 dimensions[3]. The number of dimensions (factors) used is an important issue. The original matrix of equations can be perfectly reconstructed from the SVD solution if enough independent dimensions are extracted. However, for our (and nature's) purposes this is not an advantage. Very small dimensions (small singular values[4]) represent very small, possibly locally unique components,

larger ones the components that matter most in capturing similarities and differences. One can think of the dimensions as abstract features. The features do not correspond to any characteristics nameable in words. They correspond to the foundation on which words are constructed, not to words themselves. Like the coordinates that define geographical relations, the dimensions can be rotated and scaled in any linear manner without changing anything. Dropping dimensions that don't matter is an advantage for detecting similarity. For example, a fairly common word may have been used in an unusual (or erroneous?) way a few times by some author. A learner that wants to understand a language will do better by ignoring aberrant or statistically unreliable meanings and focusing on the common core that is shared across contexts (and, thus, speakers). This is one of the effects of dropping small dimensions. More generally, dimension reduction is an inductive process that makes things more similar to each other in a well controlled manner; it is somewhat analogous to decreasing the resolution of an image by lateral integration. The great appeal of SVD for performing this function is that, in a well defined sense, it supports optimal dimension reduction. Systematically dropping dimensions from small to large retains aspects that are most characteristic and deletes aspects that are idiosyncratic or unreliable. Moreover, the analysis technique itself has discovered what "features" and combinations are most characteristic.

Evaluation of LSA's validity

A variety of quantitative simulations of human word and passage meaning, including ones in which choosing the right dimensionality has a dramatic effect, will be described later. First, some of LSA's intuitively interesting properties are illustrated. Cosine similarities (cos) are given between representative pairs of words and phrases based on a 12.6 million word corpus of general English[6].

Intuitive examples. First consider the following pairs of single words selected to illustrate various properties often exhibited by LSA semantic spaces. Random word pairs in this 300-D semantic space have cos = .02 +/- .06. (No attempt has been made to sample parts of speech or word types in a representative or random manner.)

thing – things .61

man – woman .37

husband – wife .87

woman – wife .54

man – husband .22

chemistry - physics .65

sugar - sucrose .69

sugar – sweet .42

salt - NaCl .61

cold - frigid .44

sun – star .35

sun – bright .39

sun – light .29

mouse - mice .79

doctor – physician .61

doctor - doctors .79

physician – nurse .76

man - men .41

come - came .71

go – went .71

go – going .69

going – gone .54

run – ran .57

run – runs .55

run – running .78

walk – walked .68

walk – walks .59

walk – walking .79

should - ought .51

hot – cold .48

tidy – untidy .56

good – bad .65

yes – no .52

necessary – unnecessary .47

kind – unkind .18

upwards – downwards .17

clockwise – counterclockwise .85

black – white .72

red – orange .64

red – green .47

she – her .98

he – him .93

apply – applications .42

compare – comparisons .55

comprehend - comprehension .59

detect - detectable .69

depend – independent .24

undergraduate – graduate .27

blackbird – bird .46

blackbird – black .04

Experience with many more such examples leads to the following general observations. Although no stemming or lemmatizing (reducing or transforming words to their root forms) is done, past and present verbs, and singular and plural nouns, whether regular or irregular, are usually represented as closely related, as are words related by various inflections, derivations, and compounding. Among other things, these observations raise questions about the extent of morphological analysis needed in human verbal comprehension. We find very few morphologically related words whose similarity of meaning is unrecognized by LSA despite its ignorance of morphology. Apparently inductive association from usage is sufficient in most cases. Obviously, however, it cannot explain either production or comprehension of novel meanings generated by morphological composition, which are reputedly quite prevalent in some languages (although see Tomasello, 2000, for a caution that usages, in his case grammatical structures, that appear spontaneous may actually be imitative.) The examples illustrate that semantic similarities as conceived and represented in LSA reflect world knowledge and pragmatic relations as well as lexicographic and definitional meanings.

One defect in the LSA word representation is illustrated by the antonym pairs. Antonyms are very closely related in meaning; they can be described as words that differ in only one semantic feature. Correspondingly, they are usually represented as highly similar, although further analysis can reveal that, unlike synonyms, there is a local dimension in LSA semantic space on which they differ strongly. For example, synonyms of *hot* and *cold* are all fairly close to each other but the two sets also form distinct clusters. However because antonyms are so close in LSA space, their additive effects

usually do not differentiate passages sufficiently. For example, "*A black cat is bad luck.*" and "*A black cat is good luck.*" have a cosine of .96.

Next, consider some examples of characteristic word-passage and passage-passage similarities[5]. In LSA, different senses are not separately represented; a single word-form is related to all its senses.

"*Swallow*" - "The process of taking food into the body through the mouth by eating." cos = .57

"*Swallow*" - "Small long winged songbird noted for swift graceful flight and the regularity of its migrations." cos = .30

The same pattern was found in many but not all examples studied by Landauer (in press). When word forms with multiple senses were compared with definitions of all of the senses given in WordNet (Felbaum,1998), there was a substantial, on average significant, cosine with each, even when the different meaning of the different senses was clearly reflected in relations with other words, or when WordNet definitions for the differing sentences were relatively unrelated by LSA. For example, consider the word "*fly*", for which WordNet lists 21 senses. Table 1. shows cosines between "*fly*" and words related to two senses. The two words related to each sense are closely related to each other, and the word "fly" is closely related to them all. However, the average similarity of words for one sense to those of the other, .09, is not significantly above that for a random pair. (Note that this set of relations is not possible in only two or even three dimensions.)

Table 1 here

In addition, the WordNet definitions for the two senses are both closely related to the word "fly", but the two definitions are less closely related, cos = .24, to each other.

1. "*travel through the air*" – "fly" cos = .34

2. "*two-winged insect*" cos – “fly” = .36

Over all the WordNet definitions for 21 senses of “fly”, the cosines between the entire definitions (excluding the word itself or any form thereof) and the word “*fly*” has a mean of .27, s.d. =.12. These results are typical.

In LSA, phrases sharing no words can sometimes have high similarities, while ones with most of their words in common can be entirely dissimilar.

"the radius of spheres" - "a circle's diameter" = .55

"the radius of spheres" - "the music of spheres" = .01

Correspondence with intuition is usually good for words and paragraphs, but is often poor for phrases and sentences, especially where local syntactic effects are large.

<u>About the nature of LSA word and passage representations</u>

Words and passages represented as dimension-reduced vectors in a high dimensional space have many of the empirical, intuitive, and introspective properties whose nature and explanation has concerned philosophers and psychologists. For example, a word never has a single deterministic meaning that can be fully and accurately transferred from one person to another by a definition. Rather, a definition of one word by other words can only provide a rough guide to its place in semantic space. A few contextual examples can help for a recipient who has good background knowledge, but are still insufficient. Before a word is well-known, it must be met several times (LSA simulating a high school graduate on average needs to have seen a word about eight times to get it right on a multiple choice test, although sometimes once will do), and the learner must have previously experienced tens of thousands of verbal contexts without it (below, and Landauer & Dumais, 1997). The meaning <u>effect</u> of a word is slightly different in and is changed somewhat not only by every context in which it appears, but potentially by every passage the person has ever experienced. The meaning of a word for one person is at least slightly different from its meaning to anyone else, and slightly different today from

yesterday. Wittgenstein (1953) is of course the most famous worrier about those properties.

However, LSA also offers an explanation of how two people can agree well enough to share meaning. If their language experience is sufficiently similar, the relations among words in their semantic spaces will be too. Taking word meaning to be relations of all words to all words (and all percepts, as discussed later), removes the necessity for any absolute referent or meaning for a word to have the same effect for you and me.

Passage vectors are the sum of their word vectors. They represent the gist of the passage, not its verbatim form (as noted by psychologists as far back and repeatedly as Bartlett, 1932, Bransford and Franks, 1971, and Sachs, 1967). Thus, after LSA or LSA-like processing, recall of a passage will not be word for word, it will be an attempt to convey the meaning as interpreted, that is, as represented by coding into a single high-dimensional vector in the recipient's semantic space.

These properties and others have often been taken to show that the meaning of a passage is greater than the sum of its parts. Here they emerge from a model in which the meaning <u>is</u> the sum of its word parts, but of a special kind. The LSA combining computation does not exhaust all of a passage's meaning, and can get it wrong for several reasons. Some of these reasons, such as dynamic syntactic effects, appear (but, I believe, are not yet provably) nonlinear. The analogical implication for object recognition is clear. Perhaps, much of that process is linear too, the assumption of whole greater than the sum of parts equally vulnerable to demotion.

<u>Systematic and quantitative evidence</u>

More rigorous evidence about how well LSA represents human meaning comes from simulation of human performance. For example, after training on general English, LSA matched college applicants from foreign countries on multiple-choice questions

from the Test of English as a Foreign Language. After learning from an introductory psychology textbook it passed the same multiple-choice final exams as university students. Differences in knowledge from before to after reading a technical article, and between students in different school grades, were reflected more sensitively by grades based on LSA than by grades assigned by professional readers. The following two sub-sections give more detail.

Multiple-choice vocabulary and domain knowledge tests. In all cases, LSA was first applied to a text corpus intended to be representative in size and content of the text from which the simulated humans gained most of the semantic knowledge to be simulated. In one set of tests, LSA was trained on a 12.6 million word systematic sampling of text read by American school-children,[6] then tested on multiple choice items from the Educational Testing Service Test of English as a Foreign Language (TOEFL) (Landauer and Dumais, 1997). These test questions present a target word or short phrase and ask the student to choose the one of four alternative words or phrases that is most similar in meaning. LSA's answer was determined by computing the cosine between the derived 300-vector for the target word or phrase and that for each of the alternatives and choosing the largest. LSA was correct on 64% of the 80 items, identical to the average of a large sample of students from non-English speaking countries who had applied for admission to U. S. colleges. When in error, LSA made choices correlated with the frequency of choices by students (product-moment r = .44) approximately as the average correlation between a single student and the group distribution. Importantly, when the number of dimensions was either much less or much greater than 300, the model performed much less well. At either the three dimensions of the semantic differential or the 66,000 dimensions of the original word-passage co-occurrence matrix, it got only one-fourth as many questions right.

In a second set of tests, LSA was trained on popular introductory psychology textbooks and tested with the same four-alternative multiple choice tests used for students in large classes (Landauer, Laham and Foltz, 1998). In these experiments, LSA's score was about 60%—somewhat lower than class averages but above passing level. LSA generally had most difficulty with the same kinds of items that college students do. It got more questions right that were rated easy by the test authors than ones rated of medium difficulty, and more of those rated medium than difficult. It did better on items classified as factual than conceptual. As expected, it was handicapped by questions expressed in complex sentences or containing partially irrelevant verbal content. In this case the nonmonotonicity with dimensionality was much less dramatic, but performance nevertheless decreased rather than increased after about 500-1,000 dimensions.

Essay exams. In these tests, students were asked to write short essays—varying from around 50 to 500 words over the various tests--to cover an assigned topic or to answer a posed question. The experiments have involved a wide variety of topics, including heart anatomy and physiology, neuroscience and neuropsychology, experimental psychology of learning and child development, American history, business, sociology, information technology, and others. In one case elementary school students wrote open-ended creative narrative stories constrained only by a scene-setting sentence fragment. In each case, LSA was first trained on a large sample of text from the same domain as the question. The intent is to give it text as much as possible like that from which a student writing the essay or a human evaluating it would or could have acquired the necessary knowledge of words and concepts. Each essay is represented as the vector sum of the vectors for the words it contains. Properties of these essay vectors are then used to measure the quality and quantity of knowledge conveyed by an essay, usually: (1) the semantic similarity (cosine of the angle between vectors) between the student essay and previously graded essays, and (2) the total amount of domain specific content, measured

by the essay's vector length in the semantic space for the domain. The idea is to use the way experts have scored or commented upon very similar essays to predict how they would score a new one, just as a teaching assistant may learn how to score essays by reading ones scored by the teacher. This is how the content of the essay is scored. The full essay grading system uses additional variables based primarily on other statistical language modeling techniques to reflect aspects of style, mechanics, and word order, but these measures never contributed more than 25% of the predictive variance in simulating human essay quality judgments.

In each of the experiments, two or more human experts independently rated the overall holistic quality of the knowledge reflected in each essay on a five or ten point scale. The judges were either university course instructors or professional exam readers from Educational Testing Service or similar professional testing organizations. The LSA measures have been calibrated with respect to the judges' rating scale in several different ways, but because they give nearly the same results only one will be described here[7.] In this method, each student essay is compared to a large (typically 100-500) set of essays previously scored by experts, and a subset of the most similar identified by LSA. The target essay is then assigned a content score consisting of a weighted combination of the scores for the comparison essays. In experiments, training and calibration is always performed on training data other than those used to test the relation between LSA and expert ratings.

The most notable result was that overall the LSA-based measure correlated as highly with a single human's scores as one human's scores correlated with another. On over 15 topics and a total of over 3,500 individual student essays, the LSA score were correlated 0.81 with a single average human expert, while two independent human expert scores were correlated .83 (Pearson product-moment correlation coefficient, computed on continuous LSA score values against whatever human scores were reported. It is not a

percent agreement score by grade categories.). Thus there is almost no difference between the reliability and accuracy of the LSA-based evaluation, based on its simulation of human passage meaning similarities, and that of human judges. The larger the number and variety of essay grades there were to mimic, the better the humans graders agreed with each other, and the better the training data approximated the source from which humans would have learned, the better LSA simulated the humans. All this implies that LSA and human judgments in these applications must reflect primarily the same qualities and quantities.

It is possible for machine learning technique to outperform humans, for example, because they can compare every essay to every other no matter how many. However, superior machine performance is difficult to demonstrate so long as the criterion is agreeing with human judges. If there is such a thing as a “true score” for an essay, and each human grader produces an independent noisy approximation thereto, the machine might correlate more highly with the true score than either human, and thus more highly with each human than one correlates with another. However, differences between human essay graders may be largely real differences in criteria as well as random errors in accuracy. This, together with the high reliabilities desirable and obtained in such experiments, leaves little room for demonstrating superiority of machine grading in this way. In no case has there been a $p < .05$ statistically significant advantage for the system as evaluated in this manner.

Another way to make such comparisons is to determine how well the method separates groups known to be different in the characteristics being tested. For example, the narrative essay exam was intended to be used to qualify students for promotion to a higher school grade. Therefore, we determined how well the machine scores classified students by their current grade as compared to human scores. Here, the machine was the clear winner. Measured by the pooled within-group standard deviation, the differences

between average scores of essays by students in different school grades were 50% larger by machine than human, $p < .001$. One must, of course, ask whether the machine might have achieved its superiority by sensitivity to clues—perhaps age-specific vocabulary or sheer length—that are not desirable bases of promotion. The very high correlation between machine and human scores in this case (.9) is reassuring, but a more definitive answer would require more extensive construct validity research. (More detail on machine essay grading psychometrics can be found in Landauer, Foltz and Laham, in press).

An especially interesting aspect of these results is that the machine version of verbal comprehension takes almost no account of word order; each training passage and each essay is treated as a "bag of words". Human readers presumably use syntax as well as the mere combination of words, yet they are no better at agreeing on an essay's quality. The most dramatic case was scoring creative narrative essays. One would expect order dependent syntactic factors, as in “John’s small white cat chased away the large black dog lying behind the barn” to be important in human judgments. It is possible that in some cases such as this syntactic factors were so nearly equal across individuals and groups that they contributed too little to be measured. That seems unlikely if syntax and word skills are learned or applied with any independence. In any event, for the human graders, information about relative student performance conveyed by word order must have been almost entirely redundant with information that can be inferred from the combination of words alone.

H3. A theoretical issue

These findings raise an important theoretical question. The widespread preoccupation on matters of sentential and discourse syntax in linguistics, psycholinguistics, and most natural language processing in artificial intelligence research would appear to assume that complex non-linear relations in the order of words are

necessary for adequate representation of verbal meaning. However, by any reasonable interpretation of "meaning", human judges of the knowledge content of an essay rely on meaning, and any system that can do as well as the humans using the same evidence should be considered as doing so too. It is true that superficial features of student essays that are nearly meaningless, such as the number of words, the length of sentences, the distribution of punctuation marks, capitalization, or balancing of parentheses, through correlation over individual differences in various aspects of writing ability, can generate scores that are also well-correlated with those of human graders. However, LSA uses none of these, nor any other indicator that would be little influenced by a change of meaning. It uses only the student's total choice of words for a whole essay of typically 50 to 500 words. Note that it is the vector for the total mix of words in an essay that must be right, in the sense of being correctly related to the vector for the total mixes in other essays, even though the actual words are different. Larding an essay with jargon words used out of context, for example, can sometimes make an essay score lower rather than higher, just as it sometimes does for human judges.

Other evidence

LSA has been directly compared with human verbal knowledge in several additional ways. For examples: (1) Overall LSA similarity between antonyms equaled that between synonyms in triplets sampled from an antonym/synonym dictionary, with cosines more than three standard deviations above those of randomly chosen word pairs. For antonym but not synonym pairs, a dominant dimension of difference could also be identified by computing similarities between each member of the pair and an additional set of related words from a standard thesaurus and extracting a first principal component from the intra-set similarities. (2) When people are asked to decide that a letter string is a word, they do so faster if they have just read a sentence that does not contain the word but implies a related concept. LSA mirrors this result with significant similarities and

corresponding effect sizes between the same sentences and words (Landauer & Dumais,1997). (3) Anglin (1993) had children and adults sort words varying in concept relations and parts of speech. LSA similarities correlated with the group average sorting as well as individual sorts correlated with the group average. (4) When LSA cosines were used in place of human judgments of the semantic similarity between pairs of words, virtually identical category structures were obtained with hierarchical and multidimensional scaling (Laham, 2000).

Sample applications. I view the use of cognitive models to stand in for a range of actual practical human performances as an important test of their adequacy and completeness. LSA has been used in a variety of experimental applications--including the essay scoring techniques--which were originally conceived as assessments of the model, but which have become practical applications. Here are some other examples.

(1) The technique has been used to improve automatic information retrieval, where it produces 15-30% gains in standard accuracy measures over otherwise identical methods by allowing users' queries to match documents with the desired conceptual meaning but expressed in different words (Dumais, 1994, Berry, Dumais & O'Brien, 1995). Matching queries to documents in such a way as to satisfy human searchers that the document has the semantic content they want involves an emulation of human comprehension. Surprisingly, the field of information retrieval research has never developed a technology for comparing the accuracy of machine and human performance in this task, so we do not know whether the LSA enhancement meets this objective.

(2) By training on overlapping sets of documents in multiple languages, LSA has been able to provide good retrieval when queries and documents are in different languages. The overlap need not be extremely large. Here is an example of how it works. One of two 300 dimensional semantic spaces would be derived from a year's worth of English newspaper stories, and the other from newspaper stories in Chinese for the same

year, with around a thousand stories in Chinese translated and added to the English corpus. Then the two spaces would be rotated into maximum correspondence of vectors for the subset of corresponding translated and original stories. The rest of the English stories would then be close to ones in Chinese that recount similar events, and rest of the English words close to Chinese words of similar meaning. Results tend to be somewhat noisier than those of LSA-based information retrieval on a single language. There are several reasons, among which two are of some interest. First, when one considers the different ambiguity of words and their translations, e.g. *room* and *chambre*, their relative positions in their respective semantic spaces should not be identical because not all of their "senses" (i.e., by LSA, their semantic space loci relative to other words) are the same. Second, it is often claimed that there are words or passages in one language that cannot be translated adequately into another. The LSA representation makes this intuition concrete. An "untranslatable" word is one which, when two spaces are aligned, is not near any in the other language and cannot be well approximated by the vector sum of any small number of other words. An "untranslatable" passage would likewise be one whose position is very difficult to approximate.

The other side of this coin is an hypothesis about second language learning. Human learning of a second language by immersion might go on in much the same way; inducing a separate semantic space for the second language and aligning it with the first by the overlap of a relatively small number of explicitly equivalenced passages. For the human, the equivalences could be constructed by the person's own first-language rendering of an event and that of a speaker of the second language. Such a process would make second-language learning much more rapid than first because the second borrows the structure of the semantic space of the first.

(3) LSA-based measures of the similarity of student essays on a topic to instructional texts can predict how much an individual student will learn from a particular

text (Wolfe et al., 1998, Rehder et al., 1998). The principle involved is a version of Vygotsky's zone of proximal development that we have dubbed "the Goldilocks Principle". A first-order technique finds the optimal absolute difference between the semantic content of the essay and an instructional text. This narrows choice to better candidates, but it does not distinguish texts that are optimally more sophisticated than the student from ones that are the same degree less sophisticated. A more advanced technique uses unidimensional scaling to place all the texts and essays on a common dimension. (This still doesn't specify which direction on the line is more and which less, but that is trivially determined by inspection.) Experiments estimated that using LSA to choose the optimal text for each student rather than assigning all students the overall best text (which LSA also picked correctly) increased the average amount learned by over one standard deviation (Rehder, et al. 2001).

(4) LSA-based measures of conceptual similarity between successive sentences accurately predicted differences in comprehensibility of a set of experimentally manipulated texts (Foltz, Kintsch and Landauer, 1998). The LSA method predicted empirical comprehension tests results with college students as well as the hand coding of propositional overlap used in creating the differentially comprehensible paragraphs. Prediction by literal word overlap between sentences had a near zero correlation.

(5) LSA has been used to evaluate and give diagnostic advice to sixth-grade students as they write and revise summaries of text they have read (E. Kintsch et al., 2000). Use of the system resulted in one standard deviation better summaries as measured by blind ratings, and the effect generalized to writing summaries a week later without the system's help (Steinhart, 2000)..

(6) LSA has been used to assess psychiatric status--schizophrenic or depressed patients compared to normal controls--by representing the semantic content of answers to psychiatric interview questions (Elvevåg, Fisher, Weinberger, Goldberg & Foltz,

unpublished). Accuracy was as good as those that have been reported for clinical diagnostic reliabilities of mental health professionals (e.g. Regier, et al.,1998).

Some comments on LSA representation

LSA's high-dimensional representation of meaning has intuitive appeal both psychologically and neurologically. A word has a graded degree of similarity to every other word, and it can be similar to two words that are quite dissimilar to each other. The same is true of passages. The meaning of a word or passage will be slightly different for any two people because they will have different language experience, but will be sufficiently similar that they can understand each other's utterances if their experience has been sufficiently similar. The dimensions or features that one person has used to construct a semantic space need not be the same as those used by another; they need only generate nearly the same angles between their vectors.

The pattern of dimension values (like factor loadings) that characterize a word or passage translate readily into patterns of neural activity generated either locally as synaptic conductances between neurons, or as neuronal connections between cell assemblies. Indeed, miniature LSA problems can be computed by certain kinds of unsupervised auto-associative neural network models, which compute an SVD on a single hidden layer. Dimension reduction is accomplished by converging inputs, analogous to that between retinal receptors and optic nerve fibers. While the brain surely does not use the same SVD algorithm as Berry's Linear Algebra Package (LAPack), there is no obvious reason that it can't do something equivalent using its massively parallel computational powers.

Some limitations, criticisms, and rejoinders concerning LSA

LSA as used to date has taken its input exclusively from electronic text. Obviously, most human language learners have other sources of information. They hear many more words than they read, and spoken language is considerably different from printed. They

have important perceptual inputs from the world around them and within them that accompany much of the word usage they observe. Humans also practice producing language and observing its effects on human receivers. And they get some direct tuition about word meanings. The lack of these sources of information must limit LSA's ability to model the human capability.

Grounding

An important function of language is to communicate and think about non-linguistic objects and events. LSA trained on electronic text knows about the "real world" only vicariously, by how it has been written about, perhaps somewhat akin to the visual world knowledge that a blind person has (Landau and Gleitman, 1985). We have seen that it does remarkably well with this impoverished input, much better than most people would have thought possible. This can be taken as a testimony to the power of language; language alone is able to teach a large portion of what needs to be known to mimic important parts of language and knowledge. Nevertheless, LSA surely misses, not just something, but much.

What does it miss? Some psychologists and philosophers have been especially worried by the lack of "grounding" and "embodiment" in computer models of language and thought. Grounding apparently refers to connecting language and thought to objects and events either as they "really are" or as perceived without the mediation of language. Embodiment refers to experiences concerned with states and operation of one's own body. These theorists justifiably attach special significance to these experiences and their mechanisms. The evolutionary and current adaptive success of living things is deeply concerned with maintaining bodily states in the external environment, and many of the perceptual events that inform us about them are either essentially private, unique to animal or human body and mind, pragmatically difficult to share, and/or unverbalizeable.

However, while these factors make grounding and embodiment special, and may make it more difficult, perhaps even impossible, to simulate human cognition completely with a machine, their existence in no way implies that the computational mechanisms by which they operate are different from those that deal with easier to observe or less adaptively ancient or important matters. Indeed, it is a commonplace of evolution that organic functions tend to be conserved and re-purposed rather than replaced. Moreover, there is no *a priori* reason to suppose that the mechanisms used to induce similarity between word and passage meanings is newer, less important than, or different from that used to relate the perception of external objects and internal body workings to words and thoughts. All must have evolved together from the time *Homo sapiens* had more than one word in its vocabularies.

Suppose we could get at raw perceptions and motoric intentions to encode in ASCII. We could put them into equations along with the passages of words--spoken as well as printed if we could—in whose company they do and don't appear in nature. Most would emerge from dimension-reduced SVD as very close to words; the words "headache", "fireplace", "throw" and "kiss", for example, would surely have quite high cosines with their perceptual equivalents. "Unverbalizeable" cognitions about the "real world" would be represented in semantic space as points not easily approximated by a sum of words.

LSA is, of course, incomplete as a theory of language, or even as a theory of verbal semantics. It includes no model of language production, or of the dynamic processes of comprehension. Nor does it deal with discourse and conversation conventions, or with pragmatic factors in semantics. That no current theory is more complete, and none as able to model full-scale vocabulary learning, is no excuse. We need more. However, LSA provides a good base camp for further exploration. It gives an example of an effective

computation for some important aspects of the problem and opens up paths that were previously closed by incorrect assumptions.

Some defects that LSA does and doesn't have.

Some of LSA's current incompleteness is due to technical or practical issues, not failures in principle, while others are inadequacies in need of remediation in future theory. Some examples of the former are found in Glenberg and Robinson's (2000) purported tests of LSA, which used LSA as instantiated by the University of Colorado public-access web-site tool and its general English corpus. The researchers constructed sentences in which imaginal perceptual reconstructions of real-world events were presumed to play an important role. They reported that LSA often failed to make distinctions between sentences that human subjects did. As discussed both above and later, LSA's lack of direct perceptual experience, its insensitivity to sentential syntax, and other problems as well, insure that such examples can be found. Thus, I do not doubt that LSA's representation may be faulty in this respect. However, the data that were alleged to demonstrate this particular inadequacy were badly flawed in a manner that I would like to forestall in future research. At least ten important words in Glenberg and Robinson's test sentences did not appear at all in the database. These included six words whose interpretation was the critical focus of an experiment. For example, one of their sentences was "Kenny slimed his sister." LSA read this as "Kenny his sister." Most of the missing words were inflections of words that the LSA corpus contained only in another tense (although neither "Newsweek nor Newsweeked", two of their critical words, appears at all.) LSA does not deal with generative morphological meanings, a genuine incompleteness, but not a basis for a test of this nature. In addition, recently popular usage of some of the Glenberg and Robinson words, e.g. "floppy disk" post-date the corpus. Moreover idioms and other frozen expressions were not treated as special lexical items with non-compositional meanings in the LSA analysis.

Putting aside such obvious errors, however, it is nevertheless the case that many LSA word-word and passage-passage similarities will not correspond to human intuition. LSA is dependent on the probabilistic content of the corpus it learns from, to date at most approximating the print input for just one person, not the average or range of people, and always smaller and different from the total language exposure of even any one person. And, of course, even educated humans often have the "wrong" meaning for words. Whether the frequency of errors in LSA is really, as it often appears, greater than comparable human rates, or different in principle, is hard to evaluate.

LSA performs in low human ranges on vocabulary tests, but has never been given exactly the same data to work with. Still, even with just the right data, it would remain only an approximation. For one thing, issues such as passage size and composition, how passage equations are formed, whether passages overlap, the correct pre-processing transform, and so forth, are not resolved. In visual object recognition, it is clear that wired-in neural/computational architectures upstream to memorial representation influence the process strongly. I argue that these are less important for verbal meaning representation, but not that they are non-existent.

Syntax, LSA's most important lack

The clearest in-principle problem with LSA is that word-order dependent syntactic effects on meaning are not representable in its additive combining function. Strong effects of this kind are sometimes apparent in metaphorical expressions, and in sentences with potentially ambiguous binding, attachment, quantification, predication, or negation. In these cases, errors in LSA's representations, as measured, for example, by the similarity of two sentences are quite obvious. For example *John hit Mary*, and *Mary hit John* have cosines of 1, as do *Mary did hit John* and *John did not hit Mary*. ("not" has very little meaning—a very short vector—in LSA, presumably because its effect is not additive but multiplicative and syntax dependent) These are significant errors.

However, it needs noting that all four of the above sentences tell us about a violent altercation involving John and Mary. I do not mean to make light of LSA's inadequacies, but I want to continue to emphasize the other side of the coin, how much empirical association can already account for if treated appropriately. It would be good to know how often and how much LSA goes wrong relative to corresponding rates of error, misinterpretation, or indeterminacy of human comprehension of normal discourse. Unfortunately, we do not as yet have an adequate means to answer such questions.

It is also worth noting that the possibilities of a machine learning system, even of linear ones, are not exhausted by the current LSA implementation. For example, one could relatively easily add higher order multiple-word items, both contiguous and separated, as input components, as well as other essentially associative relations of which human learning is clearly capable. It remains very much to be seen how much more can be accomplished in the same spirit as current LSA..

Let us dwell a little more on the incompleteness issue. The Glenberg and Robinson article also raises this issue. Their purpose in the research reported and the arguments presented was to compare high dimensional semantic theories such as HAL and LSA with "embodied theories of meaning". Their test paragraphs and sentences, most of which are discursively fairly complex, are all ones whose meaning depends strongly on both syntax and pragmatic knowledge about characteristics and functions of physical objects, human bodily actions involved in their use, and outcomes of those uses. They were able to compose pairs of paragraphs and sentences in such a way that there was no appreciable difference by LSA measures but obvious differences in meaning for college students. If we ignore the technical deficiencies noted above—I'm sure results like theirs could be obtained with flawless methods—the results provide a clear existence proof that LSA is incomplete as a theory of verbal meaning[8]. No argument here. If any of my presentations of LSA have given cause to believe that LSA is a to be considered a complete theory of

language and knowledge, or even lexical semantics, I regret it profoundly. LSA is a theory of (about) those things, but not of everything about them.

However, Glenberg and Robinson take their results as a general disproof of theories of the kind, saying "Because the symbols are ungrounded, they cannot, in principle, capture the meaning of novel situations." What they have shown is that LSA can fail to match human judgments of complex, syntax-dependent passages about matters that depend on perceptual, intentional, and motor experiences to which LSA has so far had no direct access. (LSA deals correctly with the meaning of novel situations as described in novel text in most of its applications.) We simply do not know whether or how well LSA or an LSA-like model would perform with Glenberg and Robinson's materials if it had learned from the same experience as University of Wisconsin college students. Therefore, it is gratuitous to conclude that it is wrong in principle from the observation that it is sometimes wrong as implemented and trained. We seek theories are about general properties and fundamental mechanisms of how things work, not about details and exceptions arising from variables not covered by theory, even if they are sometimes interesting or important.

It is also interesting to consider Glenberg and Robinson's alternative to high dimensional semantic models in contrast to LSA. They claim that meaning is based on cognition that "evolved to coordinate effective action", that the "meaning of a particular situation is a cognitive construal" that is the "meshed (i.e. coordinated) set of actions available…in [a] situation", "which depends on affordances of the situation", which in turn "are based on the relation between objects and bodily abilities." They also appeal to Barsalou's notion of "perceptual symbols", direct representations of objects that retain perceptual information. Finally, they propose that the meaning of a sentence consists of "meshing" the analogical construal of the situation with the syntax and words in a way that represents a possible action. If I understand this correctly the idea is that one can

model in the mind the possible actions that a sentence describes in a form that analogically represents the execution of a (first or second order) simulation of the intentional and perceptual event, drawing on first-hand knowledge of what actions with what objects are possible and likely. This is an appealing idea; it offers the beginnings of a way to explain the relation between some important aspects of thought, language, and action that appear to capture analog properties of the cognition of experience (See Moyer and Landauer 1967, for an early related hypothesis.)

What it does not do, however, in any sensible way, is disprove HAL and LSA. Whether comparable representational power would or would not emerge from combining perceptual and intentional experience into these models using their fundamental computational principles (see more below on object recognition), especially if temporal order and syntax were mastered, is not addressed by making this proposal. Moreover, none of the proposed components of this hypothesis have been implemented in any way and seem impossible to implement absent more explicit statement. As they stand, the use of these ideas to oppose HAL and LSA is a case of what Dennett calls an “intuition pump”, pushing the introspective mystery of a mental phenomenon to discredit a mechanistic explanation.

However, the most important aspect of this supposed debate for the purposes of the present chapter, is the issue of incompleteness. LSA is not very good at representing the meaning of passages where they depend strongly on word order dependent syntax or real-world perceptual knowledge that is not well represented in the text corpus from which it learns. And the Glenberg-Robinson-Barsalou hypotheses does not appear to apply very well to learning to represent the tens of thousands of abstract words (like *indexical*), most of which college students have met only in print. Their claim that LSA is wrong in principle because they can make sentence pairs whose relations it does not account for is roughly equivalent to a claim that a the co-ordinate system used for a map of Colorado is

in principle wrong because distances between Madison, Milwaukee and Green Bay have not been represented.

Some old arguments and their resolution

Chomsky (19635) showed that natural language often displays systematic syntactic constructions that could not be generated by word-to-word transition probabilities. This indisputable conclusion has since somehow transmogrified into a widely accepted postulate that co-occurrence cannot explain language acquisition, and thence into part of the basis for asserting the so-called "poverty of the stimulus," the belief that the information available from observation of language is insufficient for learning to use or understand it. The assertion is most firmly and often made about learning syntax, but has also been authoritatively applied to learning word meanings (Bloom, 2000; Chomsky, 1991 a, b; Gleitman, 1990, Perfetti, 1998; Pinker, 1994). Perfetti, for example, in critical commentary on a set of LSA papers (1998), after rightly pointing out LSA's in-principle failures, as listed above, and adding discourse pragmatics such as focus and rhetorical structure to the list of incompletenesses, asserts that LSA could not be considered a theory of mind just because it is based on co-occurrence. Perfetti says "Co-occurrence learning is desperately needed for the establishment of human knowledge, including knowledge about language. But more than co-occurrence is needed because of a range of human abilities that center on the representation of non co-occurring units, especially in language." The misunderstanding may be my fault. Both in the primary statement of the LSA theory (Landauer and Dumais), and in the papers reviewed by Perfetti, the dependence of LSA on co-occurrence data as input was made clear, but how the mathematics of SVD uses these data to infer "representation of non co-occurring units, especially in language", was apparently not well communicated. I hope that the derivation of LSA from SVD as a means of solving systems of simultaneous equations as presented here will help to forestall this particular objection in the future

However, the anti-learning position on verbal meaning has deeper roots than Perfetti's complaint or Glenberg, Robinson, and Barsalou's alternative views. Chomsky stated it in no uncertain terms in several places. For example in (Chomsky, 1991a), he wrote, "In the study of the lexicon, Plato's problem [the asserted fact that we know much more than experience could have taught us] arises in very sharp form, and the conclusions have to be more or less the same as elsewhere: the growth of the lexicon must be inner-directed, to a substantial extent [Plato believed we remembered knowledge from a previous life.]. Lexical items are acquired by children at an extraordinary rate, more than a dozen a day at peak periods of language growth." He goes on to point to the infrequency of explicit dictionary-like definition of words and their insufficiency for learning without a great deal of tacit prior knowledge. Moreover, he says, word meanings are "shared knowledge; children proceed in the same way, placing lexical entries in the same fixed nexus of thematic and other relations and assigning them their apparently specific properties." Therefore, he concludes, "barring miracles, this means that the concepts must be essentially available prior to experience, in something like their full intricacy. Children must be basically acquiring labels for concepts they already have…" In a companion article (1991b), Chomsky also says " It is in fact doubtful whether conditioning is any more than an artifact, an odd and not very efficient method of providing an organism with information." And, "one may ask, in fact, whether the category of learning even exists in the natural world." These were strong words, and given Chomsky's brilliant insights on other matters of linguistics and his outstanding intellectual reputation, words capable of widespread persuasion.

LSA does just what Chomsky thought impossible. It acquires linguistically and cognitively effective, shared, relationally embedded, representations of word meanings without any pre-existing specific knowledge. And it does so by learning entirely from experience. There can be no longer be any doubt that sweeping anti-association

generalizations such as Chomsky's were made too hastily, before the possibilities had been sufficiently explored, were accepted too widely and readily, and are still too persistent.

The ubiquity and tenacity of the error may relate to one of the ways in which the position has often been stated. To paraphrase: "It is impossible/difficult to imagine any way in which co-occurrence/association could account for the properties of language/syntax/word meaning." Assuming local word-to-word conditioning to be the combining function apparently shunted many minds away from thinking about other ways to use the data of experience. Of course, the failure of current LSA to account for syntax and production is fatal to its status as a complete and correct theory of language and cognition, and there may be no way to use co-occurrence data to achieve that goal. However, there is still no proof of even that at hand, no proof that a comparable method for induction of the meaning relations among syntactic patterns from their observation cannot exist. Recent work by Tomasello (2000) shows that syntactic patterns of language production develop gradually in children, at least largely as mimicry of verbatim utterances followed by generalization through iterative substitution of terms and addition of new copied patterns. This shows at least that much more is accomplished by the use of experiential data to learn syntax than has been supposed by Chomsky and followers.

Another version of the poverty of the stimulus argument comes from mathematical learnability theory. Gold (1967) and followers have shown that language, conceived as a set of deterministic rules that specify all and only a certain infinite set of word strings, cannot be learned perfectly by observing samples of the language. Informally, the proof says that because there are an unlimited number of rule sets that are consistent with all the observed instances up to now, the next sentence may violate any one currently being held. Now, of course, LSA is mute about word order, so the only rules that would be

relevant are ones that specify what words are to be used together in sentences no matter in what order.

LSA models the representation and comprehension of meaning rather than the process by which passages are produced. Nonetheless, its nature has some conceptual relevance to production, and to this proof. A language production theory in the spirit of LSA would not assume generation by deterministic rules. Instead, one could conceive of word choice as a process of finding a set of words whose vector sum approximates the vector for an idea, itself a vector in a person's semantic space. In this case, each person's partially random exposure to the world and to language samples would make location of ideas and the content of every passage somewhat different. No determinant rules are followed, only a statistical process that results in a passage that is understood approximately as intended. This is a fundamentally different conception of how language works. It does not assume that there are any such things as ideal semantic systems or underlying semantic competences distorted by performance limitations. It is, in this conception, a statistical cultural process that, coupled with a quite general individual learning mechanism, produces sufficient coincidence in the representation of relations among words and objects to support useful communication and thought. Whether the same variety of machinery can be found behind the ordering of words after or in concert with their choice remains, or course, to be seen.

What is syntax for? The combinations of words on which LSA bases its version of comprehension are not entirely devoid of grammar. Word choice includes selection of appropriate parts of speech, case, gender, tense, number, and the like. What LSA necessarily does without is local ordering of the words. To some extent our surprise at LSA's abilities may be a function of familiarity with English, a language that uses word order relatively strictly. Other languages, such as ancient Latin and modern German, are much more tolerant of variation.

Moreover, although different grammatical forms of individual words, for example *run* and *ran*, *his* and *hers*, *goose* and *geese*, are highly similar in LSA representations, they are not identical, and can have different effects in different contexts, independent of order, as a consequence of high-dimensional vector addition. These considerations do not reduce to zero the meaning-bearing effects of word order that LSA lacks. Nonetheless, the remaining role of word order in meaning representation does not seem sufficient to explain the ubiquity, complexity, and compulsory nature of syntactical conventions in languages such as English. Thus it is worth considering what other roles order-dependent syntax plays in language. Two of these are transmission accuracy and style.

Consider a message passing scheme in which no order is required for meaning. A trivial example is taking class attendance. Calling off student names in alphabetic order makes it easier for students to understand and for the teacher to note those missing. But using alphabetic order does not change the composition of the class list, its meaning; it just facilitates its transmission and use. Order functions as an error reducing code, akin to a check sum. More generally, language users will find it easier to produce the next word if they have learned conventions for utterance order; and hearers or readers will find it easier to comprehend if words come in an expected order

Clothing, body decoration, dwelling, dance, and art styles are dictated by cultural consensus that can be both highly arbitrary and strictly enforced without being deeply or intrinsically functional. The same is obviously true of linguistic syntax.

Again, this is not to say that order dependent grammar and syntax are insignificant in language, or that their explication is either an unimportant problem or one that has been even nearly solved. Finding a computational model to explain them is a major outstanding scientific challenge.

Implications to this point

I believe that the success of LSA carries important lessons for the science of learning. On the most general level, it suggests that abandoning the search for a general mechanism by which experience can be turned into knowledge of any sort is premature. It has been fashionable this decade and the last to assume that any complex appearing psychological phenomenon is only to be explained by multiple modules, “stores”, responsible brain regions, or mechanisms. Dividing a phenomenon into separate pieces, one for each thing that it can do, and assuming each to result from a separate process, can lead to progress if the analysis is correct, but it can also obscure the truth by preventing strong inferences about general mechanisms from the fact that a system does more than one thing. A different function for heart and liver may be warranted. A different biochemical process for their cellular energetics may not. The recent learning and cognition literature is replete with assertions that the idea of a general learning mechanism for all purposes is wrong. I think that conclusion may be based on the failure to discover what it is rather than there not being one. In any event, dividing the phenomena of learning into pieces by what they accomplish could only be a first step. It might push back the fundamental issue, or replace one difficult problem with several others, but the problem of finding the computations by which it all works remains unsolved.

There is also a tendency to act as if the problem of learning has been solved when it has been analyzed into separate named functions, especially if the different functions and their interaction can be separately modeled. To take a currently debated case, the fact that damage effects and activity in hippocampus and frontal cortex are different at different stages of learning is taken by some to imply two separate modules with different mechanisms. The hippocampus is said to be responsible for short term memory with rapid learning and forgetting, and for passing memories to the cortex for slow, long-

lasting storage. Different neural net models that learn quickly and slowly have been created and shown to mimic many interesting properties of learning and performance (in turn sometimes modeled by dividing the phenomenon into pieces along lines of different things it does; for example list item-recognition memory into those things for which learners can report the circumstances of learning and those they can't.) This solution may well be correct, but it also may have inhibited a computationally and physiologically truer solution. Perhaps the hippocampus is really a way station and amplifier that modulates a single cortical learning function in such a way that memories are formed quickly and fade quickly unless the hippocampus amplifies their consolidation or the events are appropriately repeated (and implicit and explicit memories are qualitatively different functions or states of the integrated process.)

Theories and models of how humans produce and comprehend language provide more egregious examples. The worst offender, in my opinion, is explanation by positing rules. It is not that rule-based thinking or behavior is in principle an impossible mechanism; all computational operations can be characterized as execution of rules. The problem is where the rules come from and how they are executed. In AI "natural language processing" the "natural" refers only to the modeled language, not the process. The rules are stated in highly advanced, culturally invented and transmitted languages, and executed in discrete steps applied to named entities in an artificial language. How the simulated human is implanted with equivalent rules and entities is either not of interest or conveniently finessed. Unfortunately, the issue is addressed only marginally better in linguistic and psycholinguistic theories where its answer is an essential goal. When rules are invoked in these disciplines, efforts are usually made to show that certain rules would or wouldn't produce or understand language the way humans do. But how the culture-dependent rules get into the mind and how they are computationally executed is still neglected. We are often told that the rules are innate, and that only a modest amount of

learning is needed to parameterize them for a given language. This could well be true. Evolution can sometimes invent organs and processes that learning can't or doesn't. However, we still want to know how the rules work and how the learning that is left to do is done. In addition, given this state of theoretical affairs--that is, absent a mechanism for their action--it is impossible to decide whether the posited rules are anything more than descriptions of the phenomenon—akin to how many petals of what color various flowers have—rather than processes that minds execute. Neural net models of language take us some distance beyond that level of understanding, primarily by proving that some portions of the process are in principle learnable or executable in some fashion (see, e.g., Christiansen and Chater, 1999; Seidenberg. 1997). Unfortunately, to date, most of these demonstrations start with inputs, and are supervised by feedback from failure and success, that require a kind of human intervention that normal human language learners do not have.

Summary of LSA's contribution to theory of language and cognition

What the LSA work has contributed to this scene is a demonstration that a system that learns from the same empirical association input as its simulated humans can acquire a very substantial portion of the human ability to deal with verbal meaning. What is important about this is not that it constitutes a complete theory of language or cognition—it falls far short of that goal--but that it demonstrates that a portion of the problem that has been long and widely believed to be beyond the power of associative learning is not. Moreover it does its job using only linear computations, albeit ones of considerable complexity. This is sufficient to call into question all the other claims about what learning and association are incapable of because all the rest have been based on the adage "it is impossible to imagine an associative mechanism that would…".

Believers in the received wisdom will object that what remains, for example syntax and non-linear logic, have not been shown vulnerable to this renewed learning-based

attack; it is only the easiest problem, merely vocabulary, that has been cracked, and that without real-world grounding, etc. The point is that LSA offers incontrovertible proof that the strong form anti-associationism, that association can't explain really important things about human cognition, is wrong. To which a reasonable further rejoinder would be that the postulate needs weakening only to exempt one aspect; that it doesn't do the rest of the job justifies skepticism that it is right about anything. Agreed. However, what is equally justified is to continue to explore computational theories of cognition based on empirical association data. Some avenues for exploration will be sketched later. It is now time to return to object recognition.

More on object recognition

Some conjectures about object recognition are suggested by the LSA research. As discussed earlier, the identity of most objects must be based on learning, just as are the meanings of words and passages. Indeed, words and passages can be thought of as physical objects whose identity is their meaning. The power of LSA to represent the meaning of any word or passage depends on its representation of all the other words and passages it has experienced. The more words it has learned, the better is its mapping of every word in semantic space. For large training corpora, experimental variation of the number of passages containing a word and the total number not containing it showed that about three fourths of the information needed to specify a word well enough to pass a multiple choice test comes from experience with passages *in which it does not occur* (Landauer and Dumais, 1994, 1995). This accounts for the fact that the daily rate of growth of reading vocabulary in middle school children per paragraph of reading is about four times the number of newly learned words that actually appear in the text they read each day.

The powerful inferential property of optimal dimension reduction for word learning depends on very large amounts of representative experience with words used in passages.

It would not do to use only a hundred randomly chosen passages, the few hundred word types they contained, and two dimensions[9]. While every word and passage could be identified, the relations among them would not accurately represent those that a literate human would perceive. The representational space on which a word's meaning is defined must be induced from empirical associations among tens of thousands of words and tens of thousands of passages (and for perceptual grounding, presumably tens of millions of percepts) before the meaning of the average word can be distinguished from others, categorized, or combined into passages, well enough for its bearer to function normally in a linguistic community.

If the linguistic analog of a visual scene or object is one or several paragraphs, then a change in wording is the linguistic analog of a change in visual view or scene. Some passages can be "viewed" from an entirely new stance, that is have all or almost all of their component words different, and still be far more similar to the original than to any other passage likely to be encountered. Their nearest neighbor will almost always be the original. The extreme sparseness of semantic space insures that most words and passages are so isolated that they act very nearly as discrete entities. For example, in a typical 300 dimensional LSA semantic space, half of the word-word cosines were below .012, that is, their meanings were essentially orthogonal, while 99% were below .16, and 99.9% were below .40.

What would a visual semantic space look like? We can, of course, only speculate. Let us suppose that the vocabulary words for vision are the outputs of retinal ganglion cells, which number about a million, and visual passages are single-saccade scenes. At two or three saccades per second 15 hours a day for 20 years there would be about 10^9 scenes. A matrix of size $10^6 * 10^9$, in this case dense, is far beyond current SVD for even the largest multiprocessor supercomputer installations, and the implied number of matrix cells is an order of magnitude greater than the usual estimate of the number of synapses

in the brain. So the brain would have to be doing a great deal of input compacting. However, something akin to incremental SVD is obviously needed for both language and vision because the capabilities of both accrue gradually throughout life. Thus we need only be interested in the size of the reduced dimension solution. Suppose that the representation of the preserved semantics of a saccadic scene takes only a small multiple of the number of dimensions needed for a passage describing it (something like a thousand words, a page full of small type). For simplicity, let's say there are 1,000 dimensions. We would thus need to keep $(10^6 + 10^9) * 10^3 = O(10^{12})$ real values, at about ten bits each, to represent every view ever seen. This is a large, but not inconceivable number. However, one estimate (Landauer, 1986) of the rate of long term memory gain for visual information based on recognition experiments would place the lifetime number of stored bits for representing scenes at only $O(10^8)$, implying a very high degree of dimension reduction by preprocessing prior to representation in memory. This also seems plausible given what we know about the early stages of visual processing. The result would be an ability to represent in long-term memory hundreds of millions of different views of the world, each by 1,000 orthogonal abstract features.

What would be the properties of such a space? We are positing that it could be modeled by an SVD solution, the similarity of two scenes computed as the cosine between their vectors. Such a space is fantastically sparse. The only scenes that would have cosines much above zero would be ones that contained regions that are highly predictive of each other. Predictiveness would come from the solutions of enormous systems of linear equations in which a scene is the sum of its preprocessed inputs. Representation is *of* not by, as Edelman puts it, the relations between scenes. A scene containing a previously seen head in a new orientation is similar to the original and very few others. The computational problem and solution are the same. What is similar to what depends on vast numbers of experiences with scenes and on the empirical

associations—correlations—that make it possible to induce a space of only 1,000 dimensions from an input of a million dimensions, such that the vectors standing for all the inputs and scenes are consistent with each other.

A number of systems for face recognition have been constructed by methods of essentially the kind proposed here, but with very much smaller training corpora and representation spaces (See Valentine, Abdi, & Otoole, 1994 for a review.) The technique even has a name, “eigenface” coding, and is usually carried out with principal components analysis (PCA) or an autoassociator neural network. Images are converted into bit strings much as I have just described, subjected to PCA, and the largest components retained. These models have been limited to images of less than hundreds of different faces, each in only a small number of views, and represented by fewer than 50 dimensions. Moreover, they often involve preprocessing to normalize location and/or the extraction or filtering out of selected input features. But even without pre-processing, they work reasonably well at identifying faces they have been trained on, even when seen in new views, or with new expressions (e.g. John Vokey, personal communication, used SVD with no pre-normalization and 20 dimensions, with very good results.). The resulting eigenvector representations can be used to code new faces in a compact form from which reconstruction yields faces that humans identify with high precision. What I propose here, then, is not a new idea in the field of image recognition. The addition is the conjecture that, as in verbal meaning representation, a very much larger and representative sample of objects and views, millions instead of hundreds, from which to learn, and substantially more dimensions, could produce qualitatively better performance.

Visual scene representation might have an advantage over verbal LSA in the fact that successive saccadic scenes are of a more stable universe. Abstracting useful dimensions from scenes whose differences often come prepackaged in monotonic changes brought about by slow view changes, smooth object movements, and multiple

saccadic images of a constant scene, should be easier than understanding language with its erratic jumps in meaning. Empirical associations over time are also simpler and more reliable in vision. That objects have permanence allows us to have object permanence; the system can learn to predict that an object that is here now will probably still be here after the next saccade or after an occlusion.

Thus, the world that vision represents would seem to have more useful constraints to take advantage of than the world of language. Presumably, that is why evolution has been able to hardwire a large portion of the dimension reduction of vision into universally useful preprocessing analyses.

Perhaps the most important point here is that if this analogy holds, then to have object and scene recognition, generalization, and categorization work well, the system needs large amounts of representative experience to construct and populate the visual semantic space. LSA may not represent two synonyms as having nearly the same meaning until it has formed a verbal semantic space based on experience with tens of thousands of other words in context. In the same way, a visual semantic space may not be able to correctly represent the similarity of two views containing the same object until it has developed a sufficient set of dimensions along which differences for scenes containing similar objects are well represented. Thus, for faces, for example, it is to be expected that invariance will not be nearly as good for rotation in the vertical plane as in the horizontal because there is much more opportunity to learn to represent the common changes in face images that are caused by horizontal rotation than by inversion. Bin Laden’s face should be easier to recognize with his head turned than upside down. On the other hand, a fighter pilot with thousands of hours of flying experience should show much more invariance in recognition of terrain objects from different views.

Let’s return for a moment to the verbal grounding problem by which researchers in computational semantics have been beleaguered. If LSA computed on a large corpus of words can infer that both "travel through the air" and "two-winged insect" are as similar to the word “fly” as synonyms are to each other, it does not seem much of a leap to believe that if coded images containing multiple foveal views of flies had been included in contextually appropriate passages, the system would have induced a representation of a fly that generalized across views and had a high cosine with the word “fly”. The system would know what flies look like, and its language for flies (and other insects, their habits, and bites, etc) would be influenced by what flies look like relative to other objects.

This conception of the computational basis of learning in visual perception does not depend on the particular mathematics of SVD. Other dimension reducing computation--wavelets or innate support vectors or mathematics yet to be invented—will undoubtedly do a better job or be more realistic vis-a-vis the nervous system. The conception is, rather, a philosophical stance on the nature of the process. In Edelman’s terms, the claim is that representation is representation of similarity, rather than representation of structures or properties. Among other things, this stance does away with the homunculus problem; no agent or executive function has to see an image in the mind, and no impossible reconstruction of the geometry of solid objects from 2-D projections is required. It yields what animals and humans need, recognition, identification, generalization, categorization, by computations on the available evidence: empirical association in space and time.

H2. Connecting verbal, geometric and physical reality

“But,” you may say, “the way the world looks to us is the way the world *is*, at least to a useful approximation, not like a list of similarity values.” Not really or not quite; or really, but in another sense. The retinal projection of the physical world does, of course, capture the geometry of seen objects, and binocular vision adds some information from

which shape can be inferred (essentially a tiny rotation in the horizontal plane to use and learn from) and collicular and cortical projections preserve it. This means that, for the kind of system postulated here, things with similar physical shapes—down to pine needles and branches, up to skyscrapers and mountains—and conglomerates thereof, are appropriately similar to each other. As a result anything we learn or inherit about the property of a kind of physical object will be represented by a vector similar to the vector for a perception of it. The mechanism is the same as that posited earlier for grounding the meaning of words. A pine tree will be perceived as solid, its trunk as round, its needles and branches as occupying spaces in geometric relation to one another.

"Oh but" you may still complain, "the world really looks like the world really is." Here, I think, we come face to face with the stubborn philosophical "qualia" problem in consciousness. We know that what we perceive is just one version of what the world is really like, just the inferences we can make from the restricted range of electromagnetic energies and their focusing on and detection by the retina. And we know that we can't prove that a pine tree looks the same to you and me, beyond what we agree on with respect to relations and identities. But we are still rightly impressed with the accuracy of perception. Almost always we can tell which of two branches is longer, bushier, greener, and verify our accuracy. We tend to forget that our absolute judgments and memories of just how long, bushy, and green are limited to three bit accuracy, that we can't recall what's on the two sides of a penny, and that the detailed content of the last saccadic scene is gone when the eyes move. Nonetheless, we retain the conscious experience of seeing the world as it is and moves. How this state of mind could be related to a vector space of 1,000 dimensions appears mysterious at best. However, it must be related to something like that because that's all there is.

I leave it at that. My mission here is not to solve the riddles of consciousness, but to suggest that a general computational method can explain both the learning of perceptual

meaning and of verbal meaning. Because vision came before language, and is often used to ground it in a different representation of reality, it would not be unlikely that the basic computational method for turning empirical association into meaning in language was adapted from that in vision.

What about trying to test this analogy by a simulation method like the one used for verbal LSA? The needed input resolution is available; digital cameras already boast half a million pixels and we could combine them. However, three things stand in the way. One is that the human, along with other animals, has additional machinery for adaptively selecting its visual input that is difficult to emulate; a mechanically centerable higher resolution fovea (the differential resolution is itself functional for such a process, but could be mimicked) and adjustable focus in service of a cognitive apparatus for directing attention. This would help it learn by the posited mechanism, as well as help it process, because it keeps related things like views of the same object in the same retinal locus, thus reducing the generalization power needed. The second is that, as noted earlier, a large part of the dimension reduction must be accomplished by wired-in preprocessing that we probably don't know enough about to implement. Finally, the matrix that we might have to feed the system to simulate human performance, a dense matrix of rank of perhaps a million, is beyond current artificial computational capability.

Nonetheless, both tests of the principle and practical applications could be achieved by applying high-dimensional representation methods to easier problems than full-scale human simulation. The best examples we have of that to date are experiments like those of Edelman, Valentine, Abdi & Otoole (1994). and Vokey (2000). What the analogy with language suggests is that to make such systems much more flexible and general they may primarily need much more empirical association data. They may need to be trained on several orders of magnitude larger and more representative samples of visual input. Rather than a few hundred selected views of isolated objects, they may need

experience with millions of images of thousands of kinds of natural scenes from which to induce the kind of high dimensional spatial structure that would make objects embedded in scenes similar to themselves despite change in context and view. With this sketch of how the ideas behind LSA might apply, apparently quite different, domain of learning, as an example, I return to consideration of more general issues.

MORE ON IMPLICATIONS

Some of the persistent problems in the philosophy and modeling of cognition seem, if not resolved, at least narrowed by these conjectures. For example, take the so-called "frame" problem, how the meaning of a passage or recognition of an object can depend on its context. The mystery appears less deep if one thinks of stimuli as combinations of millions of elements in a space of $10^{\wedge 5}$! (! meaning factorial) possible combinations. A given word or object can then have a characteristic effect on the meaning of a whole complex experience, and several words or objects can simultaneously have several independent effects, essentially by moving the vector of an experience in orthogonal directions. The effect of a given word or object on the current experience is thus always both different from that of any other word or object and different from its effect in other contexts. Similarly the problem of "the individual", which is central to many arguments about consciousness, appears less deep. A red ball is not just a red ball, it is a red ball in a context. Despite the ability to recognize it as the same with a change of context, if the perceiver keeps track of the similarity of the ball-plus-contexts in which one red ball and another have been seen, it will represent them as different, although if it loses track or the context is ambiguous it may get confused. The philosophical move is to think simple but big--very, very big--about the space for representation, rather than to think small and marvelous of individual objects.

Of course, in both word and object instances, there can be a requirement to trace the history of contexts, a matter that has not been dealt with in the models discussed here.

Following the advocated approach, one would look for the answer in associations by temporally sequential coincidence, again in the large. For the case of words, this would take the form of computations based on words-to-following-words association matrices, as is done in the HAL model of Lund and Burgess (1996), the SP model of Simon Dennis (2001), and the language models used in automatic speech recognition (e.g. Rabiner, 1969; Rosenfeld,1996), in addition to the word-to-passage equations of LSA. There is every reason to believe that empirical associations go on simultaneously at a wide spectrum of different temporal scopes, and the brain seems eminently suited to doing that in the very large. However, none of this has been cashed out in testable computational simulations, and may be too difficult to do—because of the size of system required to generate the desired properties—until there has been a great deal more expansion in computational power.

CONCLUSION

My principal goal here has been to suggest that high dimensional vector space computations based on empirical associations among very large numbers of components could be a close model of a fundamental computational basis of most learning in both verbal and perceptual domains. More powerful representational effects can be brought about by linear inductive combinations of the elements of very large vocabularies than has often been realized. Success of one such model to demonstrate many natural properties of language commonly assumed to be essentially more complex, non-linear, and/or unlearned, along with evidence and argument that similar computations may serve similar roles in object recognition, are taken to reaffirm the possibility that a single underlying associational mechanism lies behind many more special and complex appearing cognitive phenomena. Learning from vast amounts of empirical association data coupled with dimension reduction may turn out to be a technique universally used by animal and human brains. Past counter-arguments and modeling failures from Rosenblatt,

Chomsky, Minsky and Papert, and Newell, through claims for and against connectionism have been based on the empirical insufficiency of systems of small scale and the apparent nonlinearity of many cognitive phenomena. It is well-known that non-linear functions and systems can be approximated to any degree by linear systems with sufficient numbers of components or parameters. This could mean that the linear models discussed here succeed only by hammer-and-tongs approximation to real underlying mechanisms that are complexly nonlinear. However, it is equally true, mathematically, that determinate solution of large systems of nonlinear equations is anywhere from extremely difficult to impossible[10]. Many of the techniques of artificial intelligence and the posited mechanisms of rule-based theories--including ones based on logic, theorem proving, or heuristic search—to achieve realistic complexity implicitly require the solution of huge systems of nonlinear equations. Doing that has to be equally hard for biological systems as it is for mathematics. Therefore, it does not seems unlikely that nature has adopted the same trick as applied mathematicians; where the world is highly complex and non-linear, approximate it with high-dimensional additive computations on huge numbers of parameters. Fortunately, for many of the important problems animals need to solve there is plenty of data available to allow the empirical fitting of huge numbers of parameters.

This is not to claim that the biological cognitive apparatus is entirely additive down to its roots and up to its highest levels of glory. Synaptic transmissions combine nonlinearly (although possibly as an emergent function of additive combination at the molecular level), and some people do sometimes think non-monotonic logical thoughts. Linguistic syntax may be fundamentally non-linear, although I think the question is less settled than it used to appear. It would not surprise me if it turns out that the three-fourths additive, one-fourth more complex properties that have often suggested themselves in our attempts to model linguistic phenomena with LSA is close to representative. Whether that is because the underlying system is nonlinear and only partially approximated by a linear

model, or because the underlying system is built on linear foundations on top of which some local non-linear capabilities are constructed, remains to be seen. I favor the latter because it seems easier to implement and therefore more evolutionarily plausible. Even if the actual computations are basically nonlinear, as certainly is not denied by the arguments presented here, the use of a very high, but greatly reduced, dimensional embedding space based on enormous amounts of empirical association data would remain a good candidate for a computational scheme to meet the needs of cognition. Among other important properties, such a system offers to support the ubiquitous and extensive context-dependence and inductive nature of perception and language. Nonetheless, perception, language, and thought all evince phenomena that appear nonlinear, such as symbolic reasoning and hierarchical concept structures, and rather than these being functions derived out of and on top of a basic linear system, they may be symptoms of a fundamentally different scheme, perhaps, for example, one that grows very rich tree-structures rather than co-ordinate spaces. The chief drawback of a seriously non-linear version of our model is its present computational intractability. If such a method is used by the brain, it must be computable, so the problem is not impossibility, but “merely” complexity. However, discovering the brain’s mathematical tricks and matching its computational and storage capacities would be a much more daunting challenge.

Of course, these ideas are not entirely new; similar arguments used to be made by behaviorists, albeit without recourse to any adequate theory of how complex learning might work, and is still commonly made by and for connectionist modeling. To repeat, what is added here, is argument for the power of scale. Instead of scale being the enemy of learning and cognition, both in nature and for theory, so long as an appropriate lower dimensional representation can be computed, and there is sufficient data, it is a friend

References.

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

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

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

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

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

Biederman, I. (1987). Recognition by component.: A theory of human image understanding. *Psychological Review*. 94: 115-147.

Biederman, I. and Gerhardstein, P. V. (1993). Recognizing depth rotated objects: Evidence and conditions for 3D viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*. 19: 1162-1182.

Bloom, P. (2000). *How Children Learn the Meaning of Words.* Cambridge, MA. MIT Press.

Bransford, J. D. & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*. 11, 717-726.

Christiansen, M. H. & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 157-205.

Christiansen, M. H. & Chater, N. (1999b). Connectionist natural language processing: The state of the art. *Cognitive Science*, 23, 417-437.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA.: MIT Press.

Chomsky, N. (1991a). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed) *The Chomskyan Turn.* Oxford: Basil Blackwell. Pp. 26-53.

Chomsky, N. (1991b). Linguistics and adjacent fields: A personal view. In A. Kasher (Ed) *The Chomskyan Turn*. Oxford: Basil Blackwell. Pp. 3-25.

Dennis, S. Personal communications. August, 2001.

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.

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. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), *National Institute of Standards and Technology Text Retrieval Conference*, NIST special publication.

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.

Felbaum, C. (1998) Introduction. In C. Felbaum (Ed) *WordNet: An Electronic Lexical Database*. Cambridge, MA. MIT Press. pp. 1-19.

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

Foltz, P. W., Kintsch, W., & Landauer, T. K. (in press). Analysis of text coherence using Latent Semantic Analysis. *Discourse Processes*.

Gold, E. M. (1967. Language identification in the limit. *Information and control*. 10, 447-474.

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

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

Glenberg, A. M. and 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.

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

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

Landauer, T. K. (in press). Single representations of multiple meanings in Latent Semantic Analysis. In D. Gorfein (Ed) *On the Consequences of Meaning Selection.* Washington: American Psychological Association.

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

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.

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 (Ed.), *Cognitive Science Society*, Pittsburgh, PA: Lawrence Erlbaum Associates, 660-665.

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

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

Pinker, S. (1994). *The Language Instinct*, New York: Harper Collins.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (in press). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes.*

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Kauper, B., Kessler, R. C., Norquist, G. S. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders. *Archives of General Psychiatry*, 55. 109-115

Reyment, R. & Jöreskog, K. G. (1996). *Applied Factor Analysis in the Natural Sciences.* Cambridge, UK: Cambridge University Press.

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257-286.

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

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: Basic Blackwell.

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

Ch. 13. Landauer The computational basis of learning.

Author Note

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

Research and writing of this chapter were supported in part by a grant from the Army Research Institute and by contracts with Knowledge Analysis Technologies, LLC. by the Army Research Institute, the Office of Naval Research, the Air Force Research Laboratory, and other government agencies. Attributing wisdom, creativity, and discerning criticism, but not error, I thank the many colleagues who have made essential contributions to the formulation, testing and application of ideas peresented here. These include, but are not limited to George Furnas, Susan Dumais, Lynn Streeter, Peter Foltz, Michael Littman, Dian Martin, Darrell Laham, Walter Kintsch, and Bob Rehder.

Correspondence concerning this chapter should be addressed to Thomas K Landauer, Department of Psychology, University of Colorado at Boulder, Campus Box 344, Boulder, Colorado. 80309-0344, Electronic mail may be sent via Internet to landauer@psych.colorado.edu.

Footnotes

1. There are, of course, other approaches and theories for object recognition, for example the physical structure inference procedures proposed by Marr. My principal goal here being the presentation of a different way of thinking about associative learning, I do not discuss alternative object recognition theories any further. Edelman (1999) presents a good review of the field that is consistent with the present discussion.

2. SVD is a form of eigenvector/eigenvalue decomposition. The basis of factor analysis, principal components analysis, and correspondence analysis, it is also closely related to metric multi-dimensional scaling, and is a member of the class of mathematical methods sometimes called spectral analysis that also includes Fourier analysis.

3. State of the art at this writing is a newly released parallel SVD program by Michael Berry of the University of Tennessee running on a multiprocessor system with multiple gigabytes of RAM.

4. Singular value = square-root of eigenvalue.

5. The distribution of cosine values for words to passages and passages to passages is different from that for words to words. It is difficult to construct a proper representative sample for these cases because they depend on the length of the passages. However, for purposes of the examples and arguments in this chapter, it is sufficient to assume that both mean and s.d. are about twice that of word-word cosines, i.e. around .04 and .12.

6. This text corpus was generously provided by Touchtone Applied Science Associates, Newburg, NY, who developed it for data on which to base their Educators Word Frequency Guide.

7. In one interesting variant of the calibration method, only the essays themselves are used to establish the quality score. The distance between each essay and each other is subjected to unidimensional scaling, an analysis that derives the best single line on

which to locate each essay so as to maximally preserve the full set of inter-essay distances. The linear position of the essays is then taken as the content score. This procedure is analogous to a human grader who is not expert in the domain reading all of the essays, comparing each one to all the others, then ranking them from best to worst on the quantity and quality of consensual content. The procedure requires roughly three times as many essays for comparable accuracy, again much as a human might.

8. There are other defects in the Glenberg and Robinson research. (a) The subjective composition by the experimenters of passages that did not differ on LSA measures but were obviously different to college students is both highly selective, producing special examples whose representativeness is unknown, and capitalizes on the noisiness of the small-corpus dependency of the LSA. Thus, they may have used passages whose LSA representations were wrong "in principle"—LSA principle. This kind of research is more useful if words and passages are selected from natural sources by an unbiased, systematic, and objective method. (b) The suitability of sentences and word meanings, and the meaning of paraphrases used as outcome measures in their Experiment 3, were based on subjective judgments of the authors and their undergraduate research collaborators. (4) Some of the important statistical tests comparing humans and LSA made the common error of concluding a difference because one comparison was significant and the other not, instead of the correct procedure of a direct test of the difference in effects. Nonetheless, I do not think the methodological problems, separately or combined, vitiate the results; it remains clear enough that, as to be expected, LSA can often go wrong with materials of this kind.

9. Some research and knowledge engineering efforts have tried to shortcut the need for large corpus training in applying LSA, using only passages of direct interest. In most cases this has led to very poor results.

10. Recent developments in other decomposition techniques, such as Fourier wavelets, open the possibility of eventually building sufficiently constrained non-linear models for dealing with the phenomena in which we are interested. They have already enjoyed considerable success in representing visual objects for purposes of compression. Wavelet methods require ex-cathedra imposition of structure on the domain of analysis. Perhaps something of the sort is what the innate readiness to learn language and objects consists of. I stress linear systems here only because we know how to solve them at the needed scale and messiness. I certainly do not rule out the possibility that nature has discovered another way.

	insect		mosquito		soar		pilot
fly	.26		.34		.54		.58
		.61				.27	
				.09			

Table 1. In LSA a word with multiple “senses” is similar to them all, even when they are not similar to each other.

Here *Fly* has high similarity to all four words (top row), *insect* and *mosquito* are highly similar to each other, as are *soar* and *pilot* (middle row), while the average similarity of *insect* and *mosquito* on the one hand to *soar* and *pilot* on the other is quite low (bottom row).