DISCOURSE PROCESSES, 25(2&3), 363–377

The Limits of Co-Occurrence: Tools and Theories in Language Research

Charles A. Perfetti
Learning Research and Development Center
University of Pittsburgh

The development of new tools such as those described in the articles of this special issue marks an important advance in discourse research. These tools include ways to test competing models and ways to track very large language databases. The former aid the field's ability to demonstrate theoretical progress, and the latter can perform computations beyond what a human expert could do in a lifetime. There is a difference between tools and theory, however, and in this article, I point out some of the ways in which large dimensional space methods such as Latent Semantic Analysis and Hyperspace Analog to Language fall short of being plausible theories about psychological reality. I examine in-principle failures and wrong-kind failures that arise in the systems and point out the limitations of systems based exclusively on co-occurrence.

In his *Modularity of Mind*, Fodor (1983) recalls Ogden Nash's warning, "If you're called by a panther, don't anther." The point was to illustrate the advantage of a modular, highly reliable panther-recognition system over a slow and variable problem solver that, although invaluable for the higher mental processes, might not figure out whether something was a panther until it was too late. One question for this set of articles is whether the text research tools they propose can detect a panther. Their computational properties are impressive. They perform labors—counting, collapsing, rotating, dimensioning—that are beyond what a human can do in a lifetime of text processing. But I do not think they can detect a panther, although given a large enough input, they may compute panther approximations.

Before exploring pantherhood, a few general remarks are in order. First, the articles collected in this special issue represent remarkable, even exciting, progress in quantitative approaches to text. They demonstrate tools that should define the

Correspondence and requests for reprints should be sent to Charles A. Perfetti, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260.

state-of-the-art in text research for the immediate future. The range of applications appears virtually unlimited at the moment. Specific and principled failures, although possible to anticipate, are yet to be discovered. In a word, bravo.

To place these tools in context, it is useful to distinguish two components of text analysis or, more generally, language structure analysis: the analysis of the text itself and the quantitative assessment of the outcome of this analysis. The first of these—the analysis of text itself—is an attempt to reveal those internal structures of texts that are hypothesized to have some psychological reality. This has been the central preoccupation of theoretically directed research, but its quantitative aspects have been typically restricted to counting things (numbers of propositions, numbers of causal links, etc.) that are assumed to be the text embodiment of the hypothesized psychological reality. Researchers count propositions because they are assumed to be atomic building blocks in language understanding (e.g., Kintsch, 1974), and they count causal links because these are assumed to be important organizers of a readers' text understanding (e.g., Trabasso & van den Broek, 1985). This set of articles, however, reverses the usual order of things. Researchers count things—or develop algorithms that count things—and then, relying on established empirical results in the field, connect the counting to some hypothesis or empirical generalization about text processing. In either direction, the outcome of the counting and connecting implies an assessment of psychological reality.

The second component of analysis—the application of quantitative methods to the outcomes of the first kind of analysis—helps inform the assessment of this psychological reality. Thus, the first effort yields such theoretical entities as story grammars, causal networks, and propositional text bases. The second effort yields model-fitting methods, such as multiple regression, and associated least squares analyses methods, such as multidimensional scaling. When research turns its attention to this second component—assessment—the question becomes one of taking psychological reality seriously enough to test alternative ideas about its fundamental structure.

The collection of articles in this special issue contains mixes of these two research components, but the emphasis is clearly on the first—psychological reality. The article by Golden alters the mix somewhat by focusing on a general method to assess competing analyses. It demonstrates the progress to be made when quantitative model-fitting methods can be applied to a set of text data for which alternative models can make predictions. This approach, and the development of its underlying mathematics, builds a toolbox of lasting value in research. It will be a clear marker of the progress that such tools can bring when research articles reporting text recall or comprehension data will not merely demonstrate that such data are consistent with Model A, based on Psychological Reality A′, but are better explained by Model B, based on Psychological Reality B′. (It is useful to note that the value of such quantitative methods is proportional to the generality of their underlying assumptions, so that comparing Model A and Model B does not depend on measurement assumptions that are more consistent with A than B. This in general is nontrivial when the models do not share assumptions about the basic units that constitute the units over which digraphs are to be constructed.)

The other articles may be more traditional representations of the psychological reality focus, despite their avant-garde methodologies, although the article by Britton and Sorrells does mix model building with a limited kind of model testing. They assume that learning from text is (or at least includes) the learning of connectionist networks, and their comparisons are among different algorithms that distribute activation along the network. Connectionist networks have had powerful applications in other areas, and although they have had important applications in text research (e.g., Kintsch, 1988), Britton and Sorrells give the networks a distinctive central theoretical role. As with applications of these networks in other areas, it is their role in learning—in this case combining assumptions about prior states of a learner's knowledge with the content of a text—that is interesting.

Methodological sophistication may mark the coming of age of a subfield; however, it brings more than just interesting applications: It can arise fundamental questions about the psychological reality question. The main question is whether these articles have given the field some new exciting tools or whether they have proposed new text representation theories with psychological reality. To the extent that theories claim to detect or represent meaning, language, or mental representations, have they succeeded? Can they detect a panther?

Because the clearest examples occur in the group of articles on Latent Semantic Analysis (LSA), I use mainly this set of articles to illustrate the scope of these questions. However, the questions are equally relevant to the Burgess, Livesay, and Lund article, and its demonstration of Hyperspace Analog to Language (HAL). In fact, the differences between HAL and LSA, although certainly significant to the developers at the level of algorithm, are nearly invisible with respect to basic questions. Britton and Sorrells raised similar questions, although with some interesting differences. In fact, there is a remarkable coherence across the articles in their basic approach. To the extent that any one of these approaches may propose a new representational model, it shares with the others an implicit theory: Humans' discourse processes are explained by co-occurrence data in texts (and in the world).

SEMANTICS, SYNTAX, AND THE PSYCHOLOGICAL REALITY OF CO-OCCURRENCE COUNTERS

The four articles that collectively explain and apply LSA skirt an intriguing issue: What is the psychological reality of LSA? Can it detect panthers or only construct approximations to panthers? Is LSA a tool for the analysis of texts, or is it also the embodiment of a theory of human knowledge (or text understanding, or text

production, or word meaning)? Is it a theory of semantics, or is it a tool to assess various manifestations of human performance, with only an incidental, nonsystematic connection to the principles of a semantic theory? Landauer, Foltz, and Laham, in their introduction, appear to confront this issue as a matter of taste. LSA, they pointed out, can be construed either as a

> practical expedient for obtaining approximate estimates of the contextual usage substitutability of words in larger text segments and of the kinds of . . . meaning similarities among words and text segments that such relations may reflect, or . . . as a model of the computational processes and representations underlying substantial portions of the acquisition and utilization of knowledge. (p. 260)

Given these options, which one is appropriate? All the LSA articles are admirably conservative in answering this question. They report results with LSA that impressively mimic human data and then provide caveats. Landauer et al., in their conclusion section, said: "We believe that [LSA's] validity as a model or measure of human cognitive processes or their products should not be oversold" (p. 281). As the authors pointed out, the success of LSA is impressive when it simulates averages from large data sets. It is less impressive with small corpora and with individual cases. The implications of these limits merit some consideration.

Corpora Size

Corpora size is a practical problem, and if that were the limiting factor in the success of LSA applications, there would be grounds for hoping for continued improvement. It may be instructive in this respect to note the similarity between the limits of LSA and the limits of connectionist models encountered in word identification. Seidenberg and McClelland (1989) demonstrated an impressive ability of a connectionist model to pronounce words of all types, armed with orthographic input units and phonological output units but with no knowledge about how specific words are pronounced. Its performance after a few thousand training trials was comparable to that of human participants in word-naming experiments. However, Besner, Twilley, McCann, and Seergobin (1990) pointed out that the model failed to pronounce nonwords to the extent that skilled humans could, and they argued that this was an in-principle flaw in the model, reflecting the model's designed ignorance of spelling–sound correspondences, which it was left to induce from its set of trained words. Seidenberg and McClelland (1990) had a rejoinder: The training set was too small, relative to adult readers' experience with words, to allow an induction of correspondences that was powerful enough to read nonwords. In principle, the argument went, enough training on enough words could mimic human nonword ability. Whether this argument works for word reading or not (see Plaut, McClelland, Seidenberg, & Patterson, 1996, and Besner, in press, for more recent rounds of this argument), the point is its

applicability to the LSA question. Some failures to simulate human performance are not in-principle failures but, rather, arise from specific practical limitations. If the limitations cannot be overcome, then the modeling effort inherits these limitations, and the psychological reality question loses its interest as the applications function within narrow, unrealistic limits. Solving these practical limitations is a necessary condition for even asking the psychological reality question. The answer for LSA on this score seems clear: Semantic spaces can be based on indefinitely large corpora, and the larger the better.

Principled Failures

With corpora size arbitrarily large, the question becomes whether there are also principled failures. There are two places to look for an answer to such a question: First, is there something in the architecture of the modeling system that is at odds with our best information about the architecture of human cognition? Second, is there something in the modeling data that seems not just quantitatively off (which can imply a corpora size problem analogous to quantitative failures in simulations of nonword reading) but fundamentally different from the human data? It is important to note that a negative answer to the first, by itself, does not relegate the model to "mere application." Our ideas about cognitive architecture may be wrong, or the properties that emerge from the architecture may require other levels of architecture that happen to be captured by the model.

Consider the case of syntax, to which LSA is blind. On the first possibility, we may have mistaken beliefs about the role of syntax in human language processing, and a model that omits syntax will accurately reflect its minor role. On the second possibility, the role of syntax is important in human language processing, but the model escapes the consequences of this fact because it has something that implements syntactic functions statistically (and approximately). A third possibility is a special case of this implementation escape: A model may escape the consequences of its ignorance of syntax because it is given text units without syntactic variation or because it is tested by questions for which such variation is irrelevant.

This is, in effect, how LSA escapes so far. When it is given sentences, they are syntactically well-formed sentences. Thus, LSA deals with texts written and edited so that well-formed sentences are assured most of the time. What about the fact that, in many applications, LSA can be given only words, yet it still does well? This follows from the same fact: The authors of the texts that compose the corpora have written sentences from which the words come. Syntax does not matter because it is already accounted for. The test comes from including scrambled texts in the corpora and comparing results. The experiment does not have to be done: The simulation would be as good with scrambled text as with well-formed text.

Note that the authors of the LSA articles have been very clear on this characteristic of LSA. Its lack of syntax is an architectural failure in the sense defined

previously. The consequences of this failure are hidden for practical purposes and revealed only when an unnatural input is imagined or when a more syntactically relevant question is asked. As an example of the latter, imagine the outcome of a study in which we asked LSA to identify which of two authors, each with distinct styles, wrote a third text under instructions to use the vocabulary of the other author. Lexical overlap, even the layered indirect overlap used by LSA, would give the wrong answer. The same observation applies to other cognitive structures that are beyond LSA's ken—a discourse pragmatics that include discourse focus, rhetorical structure, and so on (e.g., LSA is blind to the use of sarcasm). The failures arise in the architecture and have variable manifestations that depend on the task and the input corpora.

HAL, the Burgess et al. system, also lacks syntactic knowledge. However, it has a proxy for a primitive surface syntax in its window procedure. This procedure, in effect, may provide crude approximations to word equivalence classes based on Markov-like chains. There is ample evidence that human syntactic abilities, even with a healthy dose of lexical control, do not lie at the endpoint of statistical co-occurrence approximations. It is interesting and practically useful that these approximations do capture at least a little of the patterning of sentence construction. It makes one wonder what such a system could do if it actually knew something (like syntax) to help it organize its tallying.

Wrong-Kind Failures

In testing for in-principle failure, we need to notice what kinds of failures occur. In particular, we need to notice whether they seem to be ordinary quantitative failures, hence within the range of statistical expectations and correctable by procedures that stabilize the database, or whether they are "wrong-kind" failures—failures that arouse our suspicions that they come from the foundations of the system, from properties of a system that is intrinsically different from the system of human semantics. These failures, found amid a large number of successes, are analogous to the belief that a cat made to look like a rabbit by the addition of rabbit ears is really a rabbit and will have bunnies as offspring (Keil, 1986). It looks like a rabbit, but it lacks rabbit essence so fundamentally that only the naive among us, especially very young children, succumb to the temptation to call it a rabbit.

The LSA authors have been very forthcoming, pointing out at least some of these wrong-kind failures. Landauer et al. report the success of LSA in learning synonyms as measured by the Test of English as a Foreign Language (TOEFL) multiple-choice word meaning test. LSA did as well as the average foreign students who take this test in applying for admission to U.S. colleges. To gain this success of 65% correct, LSA was loaded with 4.5 million words of text, which provided vectors for 60,000 words. However, the error pattern of LSA and that of the students was only moderately correlated ($r = .44$); that is, there

was substantial unshared variance between LSA and students when their multiple-choice answer was the wrong one. More telling may be the single example the authors provide of an LSA error: Given *physician* as a stem, LSA chose *nurse* over *doctor*. Such an error is not expected from a student but should be expected from a system that depends fundamentally on co-occurrence of words. It appears to be a wrong-kind error, although a valid judgment on this requires examination of the range of alternatives as well as a model of human errors. It is important, if one wants to explore LSA as a model rather than a tool, to study errors made in such a task more thoroughly. Studies that seek to expose the consequences of LSA on a full range of semantic performance need to pay more attention to errors than to on-average confirmation statistics.

The claim that a co-occurrence tally would generate the doctor–nurse type of error highlights the question of whether LSA is essentially a set of co-occurrence algorithms. The LSA authors generally hold that it is not, or at least not merely a co-occurrence counter. Landauer et al., again using the TOEFL data, report that the success of LSA in the vocabulary test was only 15.8% correct when there was no dimension reduction (the equivalent of first-order co-occurrence tallies in the original text database), compared with the 52.7% (corrected for chance) at the optimal dimensionality solutions. They argued from this that "the LSA dimension reduction technique captures much more than mere co-occurrence" (p. 274). This is true, in much the same way that a connectionist model captures much more than mere co-occurrence. Both are *fundamentally* co-occurrence tallying machines; both update their tallies as they experience more input, thus allowing unforeseen consequences of layered co-occurrence to emerge. All this layering makes LSA a very sophisticated cat—but still a cat, not a rabbit.[1]

Does It Matter How Co-Occurrence Is Counted?

The short answer to the question posed in this section heading is that it matters a lot for the actual success of any of these tools. It even matters for the extent to which a system will approach some elements of psychological reality. However,

[1]Consider a thought experiment: Give an expert in some domain, or perhaps a very widely read generalist, a list of words under two sets of instructions: In Task 1, the expert generates a word similar in meaning to the target. In Task 2, the expert generates a word likely to co-occur with the target in texts in the expert's domain. Although the outcomes of the two instructions may collapse on one set of word pairs, it is more likely that there will be two distinct pairings of modest overlap. In Task 2, the participant has simulated an LSA output; in Task 1, the participant has simulated a semantic equivalence output. The LSA output here corresponds to a simple co-occurrence matrix, which should be quite poor at capturing meaning similarity, according to Landauer et al. So, the exercise becomes more interesting by now giving the participant both of the original co-occurrence terms from Task 1 and asking for a third term that might be common to the first two. This exercise can be repeated up to some point of diminishing returns, gradually approximating the optimal dimension solution for simulating human-meaning judgments, without recovering actual meaning equivalence. This is what LSA does also.

any system that is based only on co-occurrence has serious limits in its ability to capture the representation of language. Nevertheless, the existence of these limits should not be taken to diminish the value of co-occurrence data in the construction of part of the system or in an approximation to some aspects of the system's output.

This point about the simultaneous limits and power of co-occurrence is important but sometimes overlooked. The contrast between co-occurrence procedures, especially connectionism, and symbolic cognitive representation systems is often viewed in exclusionary terms. This either–or dichotomy serves to give energy to the debate on the basic character of cognitive systems. Are they essentially co-occurrence systems or nonlinear symbolic representation systems (giving rise to rules, modules, or other nonlinear expressions in the architecture)? This dichotomy, interesting as it is for epistemological arguments, can mask a point that is very important for theories of representation, especially for theories that ask how representations come to get acquired: 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.[2]

To appreciate that co-occurrence and layering are fundamental in defining the properties of a text analysis system such as LSA, consider the HAL system described by Burgess et al. It is also based on counting text strings but with a difference: What gets counted is the number of words that intervene between any 2 given words within a window of text, a window of 10 words in the reported simulations. HAL's effectiveness in limited domains of language might be taken as demonstrated by the simulations reported in Burgess et al. and in other articles by this research group. It is too soon to tell whether these simulations do more than demonstrate a restricted set of interesting approximations, which would be essential, but not sufficient, in demonstrating strong hypotheses about representation or learning. Meanwhile, the demonstrations are selective. The most intriguing syntactic effects (e.g., the separation of verb forms according to semantic constraints) simply suggest that HAL captures some of the language context effects that are based on co-occurrence. The authors themselves pointed out some cases, especially sentence-level context effects, in which HAL is not effective. What is needed, eventually, in this field are tests among similarly grounded models (as illustrated by Golden's article) to merely establish the co-occurrence "finalist" in a more interesting test of models of different classes. In this context, the limitations of HAL and LSA arise from similar sources.

[2]In principle, the goal of co-occurrence modeling systems, including connectionism, is to push the co-occurrence algorithms to their limit, allowing other kinds of information (e.g., syntactic and morphological rules) only as necessary. This is not the same as believing that syntactic knowledge is unnecessary. Co-occurrence is necessary for learning but insufficient for knowledge in the domain of syntax. Syntactic representations are necessary for knowledge but insufficient for learning.

Because HAL's algorithm is based on co-occurrence, it is completely dependent on the text corpora not only for its data but also for its theory. This is a big burden.

Unlike LSA, however, which developed an algorithm and took a wait-and-see stance on its status as tool versus representational theory, HAL, like its namesake computer-turned-humanoid from Stanley Kubrick's film *2001*, seems to have had grander aspirations from the beginning: In choosing a 10-word window, "Cognitive plausibility was a constraint, and a 10-word window with decreasing co-occurrence strength seemed a reasonable way to mimic the span of what might be captured in working memory" (Burgess et al., p. 217). The consequence of this window size is probably better simulations in general, but data are actually needed on that point.[3] Whatever the empirical effects of this decision, it is an interesting implicit assumption about the relation between the human language user and the algorithm: The algorithm assigns reduced weights for intervals from 1 to 10, and then, in effect, assigns 0 for all words more than 10 words distant, whether they are in the same sentence or not. Note also that two words from the *same* sentence separated by 6 words count no more than 2 words in *different* sentences separated by 6 words. Thus, claiming that the 10-word window was chosen to reflect cognitive plausibility appears to be a claim that human working memory degrades uniformly over 10 words and then disappears. This assumption is inconsistent with the results of research on the working memory function and its dependence on sentence boundaries during reading (Goldman, Hogaboam, Bell, & Perfetti, 1980) as well as listening (Jarvella, 1971).

What is remarkable, however, is not just that a rather implausible assumption underlies the algorithm but, rather, that such a specific assumption about memory limitations suggests that HAL was intended to be a theory and not just a tool. And not a theory of representation only but apparently a theory of processing. Why base the window size on *any* estimation of memory except to reflect the possibility that humans build up their multidimensional store of meaning information by computing some analog to what HAL computes? Of course, this may be reading too much into a small part of Burgess et al.'s text, but the series of experiments and their interpretation are consistent with an ambitious goal of demonstrating the psychological reality of the representation system. Indeed, because the experiments included syntactic and lexical–syntactic questions, there is a strong implication that HAL embodies human syntactic knowledge. In fact, the experiments fall short of any demonstration that HAL can simulate basic syntactic processes. Burgess et al. are quite circumspect when discussing the problem of inferences, when they noted that they are "not arguing that these representations [based on word neighborhoods] in any way make the inference,

[3]In fact, there are hints in the results of the sentence experiments that HAL is better with short sentences than with long. Although the authors make a rather different point about this fact, it is possible that it reflects length effects in the original data matrix.

but . . . that they can provide information bearing on relevant constraints" (p. 248). A similar, perhaps even stronger, caveat would apply to syntax.[4]

THE VALUE OF APPROXIMATIONS AND TOOLS

If the foregoing criticisms suggest someone who would turn up his nose at a fine Bordeaux because he does not like wine, I hope to correct this impression. Of course, it is a matter of taste whether one accepts certain new approaches as interesting embodiments of theories or even principles, and the gist of my discussion has been to make clear why I think neither LSA nor HAL can embody a plausible theory of semantic cognition. What they can do, however, is quite exciting and cause for some celebration. They can perform. For example, in the area of text learning, LSA has predicted the extent to which prior knowledge and text difficulty together predict learning from text; in a practical application to writing, as reported in the article by Wolfe et al., it can evaluate the content of essays relative to the text base for some domain. Indeed, the success of LSA, based on a relatively modest semantic space of 36 articles and about 3,000 word types, was comparable to the success of a human grader (relative to an objective knowledge assessment) who evaluated the quality of the student essays. LSA's ability to simulate essay graders has been shown in other studies as well, so this is a generalized talent. As Wolfe et al. point out, this talent is especially remarkable because LSA knows nothing about semantic truth (so students do not get

[4]I omit Britton and Sorrells from much of the general criticism here because the psychological reality issues are a bit different in their case, despite the fact that co-occurrence counting (or correlations that reflect such counting) is at the heart of their procedure as well. They make no claims about syntax or semantics but, rather, demonstrate the predictions of essentially a learning model that already has nicely packaged concepts and links between them. The work that would be done by syntax and by the kind of semantic computation central to LSA and HAL is not part of this learning. The issues for them, as I see it, include the extent to which the constraint satisfaction model can be successful as it moves to more complex texts, as opposed to the simple clause lists that seem to have been used in some of their experiments. More problematic is how to view the process by which readers compute the representations using constraint satisfaction. The most compelling instantiations of harmony and resonance ideas have appeared in word recognition (Van Orden & Goldinger, 1994), in which the ideas of attractors, settling, and coherence have a natural appeal in a dynamic systems framework and a plausibility in the temporal properties of the co-occurrence events. Thus, orthographic, phonological, and semantic information is available virtually synchronously and rapidly settle and cohere around a word identity within a couple hundred milliseconds. However, the dynamics of perceptual events in word recognition are quite different from those in comprehension, inferencing, and the like. It is not clear how to interpret such systems in the case of text processing: as a real-time dynamics, as resultant activation patterns that survive into long-term memory, or as approximations to the products of comprehension, as derived in accord with other theories (e.g., Kintsch's, 1988, Construction–Integration model)? And there is a host of related questions for any co-occurrence weight setting procedure; for example, how does a *not* or some other negative element alter the weight of a connection that would otherwise be strongly positive? This is an intriguing approach, nevertheless, and its focus on learning marks a departure from the other articles and opens up a wide range of applications.

credit for being right); nothing about internal logic (so students do not lose credit for being inconsistent); and, of course, nothing about syntax (so students do not lose credit for not writing grammatical sentences). It appears to follow from the combination of LSA's ignorance and its success, relative to human graders, that the things it is ignorant about are redundant (highly correlated with) with the things it picks up. As Wolfe et al. explained, students who get low grades on their essays must not only say the wrong things but must "not use the right words in quite the right way" (p. 332). LSA picks up the first, whereas a human grader should pick up both. Because they are comparable in their prediction of student learning, the second is redundant.

This is a very interesting possibility, corresponding to the assumption that form and meaning tend to be interdependent in ordinary writing. Good form but poor content, on this assumption, is a condition reserved for the distinctively skilled generalist and the mythical student "BS artist" (mythical because, his or her "art" is easily detected as fraud by the domain expert essay grader). However, although I think this captures an approximately true generalization about writing—that it is generated by global and local intentions (semantics) and quickly packaged by form—I think the more interesting possibility is one overlooked by Wolfe et al.: the added quality an essay gets because of good form. Domain experts are sensitive to form; there are better and less valued ways to make arguments; there are better and less valued ways to write sentences, to mark coherence, and to segment the body of an essay. If LSA does not need the kind of knowledge about form that such values reflect, then that is interesting and must mean that the extent to which good form is present is actually predictable simply on the extent of conceptual knowledge. This is actually plausible on the assumption that the form knowledge needed is domain specific and guides the acquisition of semantic knowledge in a domain, or follows from it (the more usual assumption), or both. In short, learning about the functioning of the heart is not just about being able to write the words *aorta* and *pulmonary artery* but in learning how to link the words in their functional and anatomical roles. Getting it right, according to the essay grader, requires a semantically true, causally coherent description. Students who know the terms tend to know the links more than students who do not know the terms, so the primitive LSA method scores well. However, we do not know from the data of these essays whether there is any extra value brought to prediction by the human grader; all we know is that the simple correlations are comparable (not exactly equivalent, and slightly higher for the human graders). There is clearly the need for more work on this problem.

The practical payoff for writing—an automatic scoring of essays—is obviously high, not for actual evaluations of essays, which will remain in the hands of humans for the foreseeable future, but for research. Studies of learning from text that focus on writing by the learner are increasingly important and could certainly benefit from automated scoring. Even more promising, from an efficiency point of view, is the possibility that, where the interest lies in student learning as opposed to

student writing, one might forego essay writing altogether. Rehder et al., in their article on technical considerations of LSA, suggest that students' knowledge could be assessed by having the students generate a list of technical terms rather than write an essay, on the assumption that such a list, when placed in the domain semantic space of LSA, would reflect their knowledge as well as the essay does. To the extent this is true—and it depends on comparing knowledge correlations between LSA and various student outputs—one can imagine the results of a machine-scored vocabulary test as well as word lists and essays; this would further suggest that LSA is an approximator of human verbal knowledge and not a model of cognitive processes.

LSA also does nicely at assessing text coherence, as Foltz, Kintsch, and Landauer demonstrate. What strikes me as especially useful in these demonstrations, as with the essay grading, is that coherence can be assessed with sentences instead of propositions. This is a gain in both convenience and tractability because sentences are well defined in texts, whereas finding propositions is both labor intensive and a bit arbitrary. Perhaps more interesting than labor saving and nonarbitrariness is the semantic basis for coherence. The value of argument overlap as an indicator of coherence has been verified very clearly in research. Argument overlap not only satisfies minimum requirements for cohesion but is a powerful predictor of text-learning performance, when coupled with assumptions about the need to link text segments in working memory. The tour de force of LSA here is to mimic the proposition-based assessment of coherence without propositions and without argument overlap. It is a matter of deeper semantic coherence, the kind of "is-this-sentence-somehow-related-to-the-meaning-of-the-previous-sentence?" that LSA picks out on the basis of its multilayered overlap assessment.

It is interesting that in one of the simulations reported in Foltz et al., the multilayered property of LSA—which is critical to its potential for a cognitively interesting model—was unnecessary. Only simple lexical overlap was needed to simulate the coherence results of Britton and Gulgoz (1991), much less than what LSA offers. In this case, LSA semantics were redundant. However, in texts that were revised more heuristically, using methods other than argument overlap, LSA demonstrated that its sensitivity to broader semantic structures predicted text coherence better than word overlap. This will be good news for those who believe that the kind of coherence captured by argument overlap is only at the margins of "real" coherence, assumed to be carried by more ephemeral, or at least less countable, characteristics than explicit anaphora. Now, one can count those uncountable, implicit, semantic events, even if they are ultimately based on co-occurrence.

Finally, returning to the larger picture, notice that LSA predicts a variety of text-processing results. In addition to the few I discuss here, consider that research has demonstrated predictive effectiveness in query-document relevance judgments, subject matter multiple-choice responses, context effects in lexical priming, and word-sorting performance. Added to text coherence, text learning as a function of knowledge and text difficulty, and vocabulary—the ones I discuss—

this is quite a range. The fact that simulations can be successful across such a broad sampling of human language use is more than impressive. It should trigger a suspicion that LSA cannot be a model in the usual sense. It does too much. Rather than explain performance fully in one or two domains, it explains performance incompletely in many domains. The parallel with connectionism is striking on this point. Each is less a model of human language processing than a powerful implementation of some important principles. In the case of connectionism, the principle is co-occurrence learning and the unforeseen consequences of that learning that make it more powerful than input–output associationism. In the case of LSA, the principle is the same: co-occurrence learning applied in iterations that produce multiple layers, analogous in some ways, despite important differences, to the hidden units of a connectionist network.

Thus, co-occurrence by itself is not the only principle, and it may not even be the most important one. The vast dimensional spaces that are developed (through singular value decomposition algorithms) in its successful applications are equally important. And what principle, if any, underlies this vast dimensionality? In my view, a useful way to think about this is that LSA dimensionality—what I have referred to as its multilayering—captures the coherence of human written discourse around a given topic. Words co-occur. Some words are more distinctive (developing greater weights) because they occur more in this topic than in that one. A system that reflects the first, second, and *n*th order co-occurrences and also reflects the weighting of concepts (among other things) is able to capture an abstract semantic space that may be nearly as "meaningful" as a semantic space organized by a set of semantic concepts (even primitives). The LSA space gains semantic power because it reflects the expressed knowledge—or, more generally, the expressed descriptions—that are applied to domains.[5]

[5]It seems to be important that these descriptions be bounded, restricted around a domain. Their power lies in part through their ability to organize a lexicon around *X* rather than around *X, Y,* and *Z.* It is instructive to imagine the consequences of a truly random and indefinitely large sample of English language texts. The statistics for a large unselected sample should approach the kind of type and token counts that are obtained from large sample corpora generally. The ability to discriminate topic importance should diminish because the universe of (nearly) all texts will include all the topics of human discourse. Rehder et al. make the point that, in their study, only articles about the heart were included in the LSA space. Off-topic words could not, therefore, affect the vector representations, including length (strength or importance). Thus, in effect, an approach that gains power in domain-sensitive modeling is to load the semantic space with material targeted for a specific domain. The implication seems to be that LSA cannot be expected to do well in a domain if it has too much off-topic knowledge. In turn, this implies a stance on the question of general knowledge and domain performance, that complete knowledge might not lead to high-domain learning. The oddity here, unlike the case for which superficially similar arguments are sometimes made (viz., that domain knowledge is necessary for domain learning), is that a full representation of all English texts includes the representation of all domain knowledge and thus is highly and equally expert in all domains. Instead of allowing the learner who has the whole world of discourse as a knowledge base to learn anything, the technical discussions of LSA in these articles imply the whole-world learner will not do well on anything with significant domain context.

HAL holds similar promise as a tool, and, indeed, it has already been applied to a large number of problems in language. Some researchers who contemplate applying one of these systems to a specific problem may have a choice: HAL or LSA. (Other applications seem to suggest one rather than the other.) We should hope that some problems might be addressed by both systems, using a single large corpus, to gain more information on the strengths and weaknesses of each in a given application.

CONCLUSIONS

The development of computational procedures for the analysis of texts and for the evaluation of text-learning hypotheses is an indicator that text research has come of age. Although the field, for some time, has had computational models of text processing, what is new here is the convergence of a variety of theoretical text-processing issues with computational methods of analysis and evaluation. In this article, I focused primarily on the methods of high-dimensional space, especially LSA (but also HAL), because this approach is generating so many specific applications. The widespread applicability of LSA-type high-dimensional spatial representations—derived from large databases—augurs a development that could become for text processing what connectionist modeling has become for cognition as a whole.

Is it a tool and a theory, or just a tool? Something so general that it is indifferent to a wide variety of structural properties that are definitional for cognition is a poor candidate for a model of cognition processes within a specific domain. The same probably holds for HAL. For something to be useful, however, it is not necessary that it detect panthers. It can organize knowledge (a database) in such a way as to allow all sorts of other things to be detected.

Language data as contained in all kinds of texts are needed both to document language use and to inform understanding of language processes. However, the data are intractable for human users beyond the most superficial and quantitatively limited analyses. These new systems can track the data; compute their various transformations; and express the underlying order that, in some sense, must be collectively represented in the language users who produced the original data. These are accomplishments of tremendous potential for progress in research.

Recognizing LSA (or HAL) for what is—a reiterative application of co-occurrence statistics into high-dimensionality space—will help researchers apply it to the problems for which it will have high value. These systems do not need the additional burden of being putative models of cognition.

REFERENCES

Besner, D. (in press). Basic processes in reading: Multiple routines in localist and connectionist models. In P. A. McMullen & R. M. Klein (Eds.), *Converging methods for understanding reading and dyslexia.* Cambridge, MA: MIT Press.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review, 97,* 432–446.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Experimental Psychology, 83,* 329–345.

Fodor, J. A. (1983). *Modularity of mind: An essay in faculty psychology.* Cambridge, MA: MIT Press.

Goldman, S. R., Hogaboam, T. W., Bell, L. C., & Perfetti, C. A. (1980). Short-term retention of discourse during reading. *Journal of Educational Psychology, 72,* 647–655.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior, 10,* 409–416.

Keil, F. C. (1986). The acquisition of natural kind and artifact terms. In W. Demopoulos & A. Marros (Eds.), *Language learning and concept acquisition* (pp. 133–153). Norwood, NJ: Ablex.

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

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

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, development model of word recognition and naming. *Psychological Review, 96,* 523–568.

Seidenberg, M. S., & McClelland, J. L. (1990). More words but still no lexicon: Reply to Besner et al. (1990). *Psychological Review, 97,* 447–452.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612–630.

Van Orden, G. C., & Goldinger, S. D. (1994). The interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1269–1291.