Behavior Research Methods, Instruments, & Computers
2003, 35 (2), 244-250

Computerizing reading training: Evaluation of a latent semantic analysis space for science text

CHRISTOPHER A. KURBY, KATJA WIEMER-HASTINGS, NAGASAI GANDURI, JOSEPH P. MAGLIANO, and KEITH K. MILLIS
Northern Illinois University, DeKalb, Illinois

and

DANIELLE S. McNAMARA
University of Memphis, Memphis, Tennessee

The effectiveness of a domain-specific latent semantic analysis (LSA) in assessing reading strategies was examined. Students were given self-explanation reading training (SERT) and asked to think aloud after each sentence in a science text. Novice and expert human raters and two LSA spaces (general reading, science) rated the similarity of each think-aloud protocol to benchmarks representing three different reading strategies (minimal, local, and global). The science LSA space correlated highly with human judgments, and more highly than did the general reading space. Also, cosines from the science LSA spaces can distinguish between different levels of semantic similarity, but may have trouble in distinguishing local processing protocols. Thus, a domain-specific LSA space is advantageous regardless of the size of the space. The results are discussed in the context of applying the science LSA to a computer-based version of SERT that gives online feedback based on LSA cosines.

According to the constructionist model of text processing, understanding written text often requires the generation of inferences (Graesser, Singer, & Trabasso, 1994). People have been shown to strategically make inferences while reading (Magliano, Trabasso, & Graesser, 1999). These inferences may be in the form of predictions, bridges, logical statements, elaborations, and metacognitive statements (Magliano, Wiemer-Hastings, Millis, Muñoz, & McNamara, 2002). Generating particular kinds of inferences, particularly while self-explaining difficult science text, helps readers to better understand the content. People who self-explain texts in a think-aloud format improve their comprehension and build better mental models of the material (Chi & Bassok, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994; Magliano et al., 1999; McNamara, 2003; McNamara & Scott, 1999; Trabasso & Magliano, 1996).

For instance, when reading a text about thunderstorm development, a reader may encounter new concepts or processes. In effective self-explanations, the reader connects the new information to his or her world knowledge or prior text information. For example, the reader might explain the sentence "One part of the cloud develops a downdraft" as "One part of the cloud begins to sink. It will probably begin to rain shortly, because it will not be able to hold any more precipitation." This explanation paraphrases the current sentence, and brings in world knowledge to explain what is going on and to predict what might happen next. This type of self-explanation helps readers build their knowledge related to the sentence topic and generally facilitates the understanding of the text (Cote & Goldman, 1999; Magliano et al., 2002; McNamara & Scott, 1999; Millis, Magliano, Wiemer-Hastings, & McNamara, 2001).

McNamara has developed a reading training technique called self-explanation reading training (SERT; McNamara, 2003; McNamara & Scott, 1999). SERT is designed to improve the process of self-explanation by teaching students to use a variety of reading strategies and make inferences while self-explaining text. SERT not only improves students' quality of self-explanations and text comprehension, but it has also been shown to improve students' course grades (McNamara & Scott, 1999). During training sessions, students are taught in a classroom setting about the different types of inferences emphasized in SERT, which are predictions, bridges, logical/common-sense statements, elaborations, and metacognitive statements that represent comprehension monitoring. They are also taught how to paraphrase a sentence. The next phase of SERT training is called demonstration. The instructor plays a videotape of a student reading a text aloud and thinking aloud using SERT strategies. The students are asked to identify the various strategies used in a subset of the explanations. During the practice phase, they practice in pairs

This research was supported by National Science Foundation grant IERI #0241144 to D. S. M. The authors thank Nicole Gurt and Jeremy Omerod for their assistance with the ratings of test protocols. Please address all correspondence regarding this article to C. A. Kurby, Department of Psychology, Northern Illinois University, DeKalb, IL 60115 (e-mail: ckurby@niu.edu).

using SERT while reading science texts (see McNamara & Scott, 1999, for a detailed description of SERT).

We are currently developing a computerized version of this training technique that can be employed in schools to help students read scientific texts more effectively (Levinstein, McNamara, Boonthum, Pillaraisetti, & Yadivalli, in press; Magliano et al., 2002; Millis et al., 2001). A computerized trainer faces a key difficulty of automatically judging which reading strategies a reader is using during self-explanation. This is crucial if it is to determine whether or not the reader is effectively using self-explanation.

One method of automatically assessing self-explanation protocols on the computer is the use of latent semantic analysis (LSA; Landauer & Dumais, 1997; see Magliano et al., 2002, for an initial evaluation of LSA in SERT). LSA determines the similarity between words by the frequency of their co-occurrence in a large corpus of text. It does this by (1) creating a word-by-document matrix of co-occurrence frequencies of all words in a corpus and (2) condensing the matrix using singular-value decomposition. The result is a high-dimensional space in which each unit of text is represented by a vector. To determine the similarity between text units, LSA calculates the cosine between their LSA vectors. The cosine generated is a measure of semantic similarity and varies from 0 to 1. A higher cosine indicates that two units of text are more semantically similar.

The main goal of this paper is to assess if LSA reliably correlates with human judgments of similarity between think-aloud protocols and standard protocols representing reading strategies. To be useful in SERT, LSA needs to be able to discriminate between different reading strategies and to identify a strategy that human raters detect in a student protocol. There is research that suggests that LSA is a good technique to use in the SERT trainer. For example, LSA has been found to correlate well with expert graders in grading essay papers (Foltz, Laham, & Landauer, 1999), and it is generally consistent with human judgments of similarity (Landauer & Dumais, 1997; Landauer, Laham, Rehder, & Schreiner, 1997; Magliano et al., 2002). In addition, LSA has been successfully used within a computer literacy tutor called AutoTutor (Graesser et al., 2000; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999). AutoTutor presents questions to students, evaluates students' answers, and gives appropriate feedback. In AutoTutor, LSA computes the semantic overlap between current student contributions and stored ideal answers or bad answers to determine their correctness. Feedback is determined, to a large extent, by LSA cosines. Choosing appropriate feedback is also a goal of the computerized SERT trainer.

Magliano et al. (2002) examined whether LSA correlates with human judgments of reading strategy. After SERT was administered, each participant's think-aloud protocol was divided into clauses and coded according to Cote and Goldman's (1999) three reading strategies (i.e., minimalist, sentence processing, and knowledge building). These three strategies were adapted for the present context as *minimal*, *local processing*, and *global processing*, respectively, which distinguish between the three sources of knowledge that readers use to make sense of the current sentence. Hence, the sources of knowledge identified by human raters for each clause were the current sentence, prior text information, and thematic or world knowledge. These distinctions are important, because enabling the SERT trainer to be sensitive to the source from which a reader is drawing allows it to automatically judge which reading strategy the reader is using. Thus, following coding, each protocol was compared against three separate benchmarks (comparison sets) using LSA. The benchmarks were representative examples of specific reading strategies and were based on the three predetermined sources listed above. Specifically, a minimal response is just a paraphrase of the current sentence or a vague comment (e.g., "good"). A local processing response may contain a paraphrase and a local bridge or an elaboration of the sentence that was just read. This type of response may clarify the sentence for the reader, but it does not connect the sentence to the main theme of the text. A global processing response connects the sentence just read to the main theme of the text, to the reader's world knowledge, and to prior text information. This response usually involves the use of multiple SERT strategies.

Magliano et al. (2002) found that LSA corresponded with human judgments in determining which protocol is from which source and that participants who received SERT produced more global processing responses. This result suggests that LSA can be employed effectively to perform the challenging task of automatically judging participants' reading strategies. However, this evaluation was made using the LSA general reading corpus of the online LSA website (http://lsa.colorado.edu). We have since developed an LSA space that is specifically designated to handle science texts. The question addressed in this paper concerns the extent to which this new LSA space can perform the task of distinguishing between reading strategies. Principally, it seems likely that a domain-specific LSA space would perform better than the LSA space based on the general reading corpus. This is because it contains high frequencies of the words that constitute key technical terms in the target texts. Specifically, the LSA space was trained on texts from earth science, physical science, and biology to deal with the scientific texts for which SERT is used. This aspect is important, because there is empirical data suggesting that LSA performs best when it is trained in a content area similar to that of the material to be analyzed (Shapiro & McNamara, 2000).

METHOD

The science LSA database built for this project contains book chapters and articles on the topics of earth science, physics, chemistry, health, and biology. Texts were taken from eight textbooks, websites, online encyclopedias, and a CD-ROM database. Altogether, there were 273 documents with a total of 849,060 words. One document corresponds to roughly one textbook chapter or online article and contains an average of 3,073 words. The range of words

across all texts in the corpus is 23,476. The corpus contained documents pertaining to general topics, such as weather cycles and thunderstorm development, as well as documents concerning specific topics, such as types of heavy storms. The general texts were typically longer than the specific texts.

Evaluating the Performance of the Science LSA Corpus

We evaluated the performance of our LSA spaces by comparing the cosines to human ratings and to the general reading space at the University of Colorado website. Cosines and human similarity ratings were obtained for pairs of student protocols and semantic benchmarks. Table 1 shows an example sentence and the benchmarks to which the following SERT protocol was compared: "Long ago coal was recognized where ancient swamps once were located."

The goal of the following analyses was to identify the LSA space that provided the highest match with the human ratings. In total, 10 science LSA corpora were compared. We started with a 100% text corpus, from which other corpora were constructed by removal of a certain amount of text that was either specific or general in nature. This resulted in the original 100% corpus, an 80% specific corpus, an 80% general corpus, a 60% specific corpus, and a 60% general corpus. For instance, the 80% specific corpus is the 100% corpus minus 20% of its general texts. Each of these 5 corpora was formatted into paragraphs and sentences, forming a total of 10 corpora. Each corpus was trained on 50 to 450 dimensionalities. If LSA reliably makes reasonable judgments, then its output should be correlated with the human judgments. Also, if this specific science database is better suited for judgments of science texts, its correlations with human ratings should be higher than those of the general reading corpus (henceforth, LSA_GR).

The SERT protocols used for this evaluation were taken from Magliano et al. (2002). The protocols were from 90 students and varied on processing levels. Protocols reflected minimal, local processing, and global processing reading styles, depending on the extent to which the students used active reasoning (Cote & Goldman, 1999). They were hand coded by two experts and two novices according to how similar they were to five benchmarks on a 6-point scale. Benchmarks 1, 2, and 3 included selected content words representing the current sentence, prior text, and world knowledge, respectively, and Benchmarks 4 and 5 were each a prototypical protocol. The expert raters were text researchers centrally involved in this project. The novice raters were students trained to make judgments with only minor exposure to the project. To evaluate the science LSA spaces, we compared their cosines against the human ratings and against the cosines obtained from LSA_GR.

RESULTS

Two sets of analyses were conducted to evaluate our science LSA spaces. In the first set of analyses, the agreement between LSA cosines and human judgments of similarity for each protocol and benchmark was examined. The difference in cosines between different levels of semantic similarity was assessed as well. We also assessed whether cosines would differ significantly between levels of rated similarity. In the second set of analyses, we investigated whether the highest LSA cosine is actually obtained with the benchmark that represents the strategy used (as determined by human classification). For example, the LSA cosines for the current sentence (CS) benchmark should be highest for minimal protocols, the cosines for the prior text (PT) benchmark should be highest for local processing protocols, and the cosines for the world knowledge (WK) benchmark should be highest for the global processing protocols. The second set of analyses also tested LSA's ability to match human judgments of processing level.

Agreement Between LSA Cosines and Human Judgments

Correlations were calculated between the averaged ratings of similarity from the two experts and the averaged ratings from the two novices, LSA_GR, and our science LSA spaces. Correlations between LSA cosines and both expert and novice human ratings of similarity can best be conceptualized by comparing them with the correlations between the human raters themselves. This correlation provides a baseline with which the performance of LSA can be compared. All correlations between the human raters were significant ($p < .001$); the correlation between the experts was .79, that between the novices was .74, and the correlations between the experts and the novices ranged from .69 to .80.

First, we will examine the influence of each factor on the performance of the science LSA spaces in terms of correlations with averaged human expert and novice ratings. This analysis identifies factors that are important for the science LSA spaces to perform well. We will then compare the performance of the science LSA spaces with that of LSA_GR to assess the viability of constructing a domain-specific corpus.

The LSA science spaces correlated with both average expert ratings ($M = 0.71$) and novice ratings ($M = 0.64$) more highly when they were formatted in paragraphs than when they were formatted in sentences [experts: $M = 0.64$, $t(46) = 5.37$, $p < .001$; novices: $M = 0.60$, $t(46) = 3.05$, $p < .001$]. Because of this difference, further analyses of the science LSA spaces will be conducted on data

Table 1
An Example Sentence and Benchmarks Representing the Current Sentence (CS), Prior Text (PT), and World Knowledge (WK)

Current Sentence	Benchmark
It was recognized long ago that coal accumulated where ancient swamps once were located.	CS: recognized, long ago, accumulated, ancient, swamps, located, fossil, coal
	PT: forty, decomposed, matter, rock, sedimentary, inorganic, less, percent
	WK: formed, place, found, used, dead, discovered, made, organic, time, area, ideal, many, moist, old, reasons

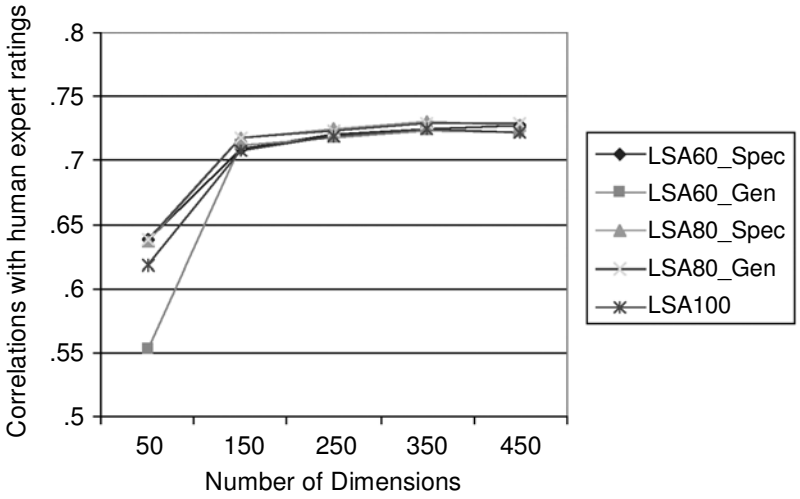

Figure 1. Correlations of each LSA corpus (formatted in paragraphs) with averaged expert human ratings across each number of dimensions.

from the corpora that are formatted in paragraphs. The cosines of the science LSA spaces, formatted into paragraphs, correlated very highly with expert human ratings (for all correlations, $p < .001$). The correlation with the average expert ratings is an important test of LSA, because the tutor is meant to evaluate reader strategies as an expert would. Figure 1 shows that the correlations get larger as the number of dimensions increases and peak at 350 dimensions. No differences in correlations were found between specific and general formatting (specific: .67 with experts, .63 with novices; general: .67 with experts, .62 with novices), and no differences were found between the different percentages of text (60%: .67 with experts, .62 with novices; 80%: .68 with experts, .62 with novices; 100%: .68 with experts, .62 with novices). These results suggest that the different style of science text (specific vs. general) and the amount of text (60% and up) do not have an effect on LSA's ability to make similarity judgments comparable to those made by humans.

In comparison with correlations given by the science corpora, LSA_GR returned cosines that correlated less highly with both averaged expert and novice ratings (experts, $r = .55$, $p < .001$; novices, $r = .51$, $p = .001$). This result suggests that the science LSA space can reliably estimate the similarity between a given student's thought and benchmarks that represent different reader strategies, and that the designated space is better matched to the task than the general reading space is.

We were also interested in whether the cosines produced for each level of semantic relatedness differed significantly from one another. This was done by comparing the cosines obtained on test protocols that had been rated as little related, somewhat related, and highly related to the benchmarks. To this end, the rating scale used for the human ratings was divided into three parts: unrelated (scale points 1–2), somewhat related (3–4), and related (5–6). A few test protocols were omitted from this analysis because of missing ratings. There were 112 test protocols for the unrelated group, 151 protocols for the somewhat-related group, and 163 protocols for the related group. Our goal was to determine what level of correlation with human ratings was needed for LSA to discriminate between different levels of semantic similarity. Thus, we tested the performance of two corpora: one that yielded the highest correlations with human ratings and one that yielded the lowest. The first was an 80% corpus that was trained on 350 dimensions and formatted into paragraphs. This corpus was used because results from the correlational analysis suggested that the paragraph-formatted specific corpus and the general one trained on 350 dimensions were the best performers. The second was the 100% corpus trained on 50 dimensions and formatted into sentences. This corpus was chosen because it had the lowest correlation with human ratings. Mean LSA cosines for test protocols in these three areas of similarity, as judged by the expert raters, were significantly different among the three groups of semantic similarity (unrelated, somewhat related, and related) for the 80% corpus, as tested by an analysis of variance (ANOVA) [$F(2,426) = 170.48$, $MS_e = 0.04$, $p < .001$]. The mean cosines plus standard deviations (*SD*) for unrelated, somewhat related, and related similarity groups were $M = 0.14$ (0.02), $M = 0.36$ (0.02), and $M = 0.60$ (0.02), respectively. Mean LSA cosines differed significantly among the three groups for the 100% corpus as well [$F(2,426) = 72.58$, $MS_e = 0.04$, $p < .001$.] The mean cosines for unrelated, somewhat related, and related similarity groups were $M = 0.38$ (0.02), $M = 0.55$ (0.02), and $M = 0.71$ (0.02), respectively. Post hoc Scheffé analyses indicated that there were significant differences in cosine size between all levels of semantic similarity for both corpora.

In summary, these analyses show that each corpus produced cosines that are significantly different from each other with respect to each level of semantic similarity.

Thus, it is quite likely that all of the corpora distinguish between different levels of semantic similarity as judged by human raters. This finding is important because it demonstrates that each of our science LSA corpora can be implemented in detecting differences in semantic similarity between student think-aloud protocols and benchmarks that represent reading strategies.

The science LSA outperformed LSA_GR in this initial analysis. Correlations of human ratings with our designated corpus were consistently higher, with an overall correlation with averaged expert ratings of $r = .72$ for the science LSA and $r = .55$ for LSA_GR (both ps $< .01$). Thus, the science LSA space stands up to this comparison favorably.

LSA's Ability to Distinguish Between Strategies Using Benchmarks

In the second set of analyses, we investigated whether the highest LSA cosine is actually obtained with the benchmark that represents the strategy a reader is using (as determined by human classification). The cosines relating each student protocol to each strategy benchmark using LSA_GR in paragraph formatting were analyzed first. We would expect to see cosines decrease for CS as text-processing level increases. For instance, the cosines for CS should be highest for minimal responses and decrease as text processing becomes more global. Conversely, cosines for WK should increase as responses become more global. That is, the cosines for WK should be lowest for minimal responses and highest for global responses. We also expected cosines for PT to increase as responses move from minimal to local processing and then to decrease for global processing responses. In other words, PT cosines should be highest for local processing responses and taper off for minimal and global processing responses. In summary, LSA cosines should be highest for each benchmark that corresponds to the reading strategy that a reader is using. Thus, we expected an interaction between text processing level and type of benchmark.

The data were analyzed using a 3 (benchmark: CS vs. PT vs. WK) $\times$ 3 (processing level: minimal vs. local vs. global) univariate ANOVA. The interaction between text-processing level and type of benchmark was significant [$F(4,261) = 4.17$, $MS_e = 0.02$, $p = .003$]. This result shows that cosines decreased for CS as text-processing level increased, whereas cosines for WK increased. However, the cosines for PT were relatively flat, so that the cosines did not peak at local processing and then taper off at each end. The same ANOVA was conducted on each of the paragraph-formatted science LSA corpora and revealed a significant interaction between text-processing level and type of benchmark for all of them and the same pattern of cosines for PT. These patterns are inconsistent with our predictions. That is, these results indicate that discriminating local processing from other processing levels is particularly difficult for LSA.

Chi-square analyses were performed to test if LSA's difficulty in producing high cosines for local processing responses is evident in a difficulty in matching human judgments of that processing level. For example, it may be possible that local processing responses are misclassified as another reading strategy. The LSA spaces used for these analyses were the LSA_GR corpus and the same 80% space used in the first set of analyses (presented above) in order to get a good comparison of LSA's performance between these two corpora. The chi-square analyses for these corpora confirms the initial finding that local processing protocols are difficult for LSA to detect. Only the cosine-based protocol classifications obtained with LSA_GR produced a significant match with the human codings [$\chi^2(4) = 14.05, p < .01$]. The science LSA space produced a nonsignificant match. As can be seen in Table 2, the difficulty for both spaces was the classification of test protocols coded at the local processing level (between minimal and global processing). The numbers indicate that the percentages of minimal, local, and global processing codings by humans were matched by the LSA cosines for each processing level. For example, only 23% of the local processing protocols were classified correctly by the LSA_GR space, and 32% were classified correctly by the 80% LSA space. These protocols were often misclassified as minimal because they draw on intertextual information. Excluding the cases of the local processing level from analysis brings the chi-square values of the science LSA space into a range of at least marginal significance [$\chi^2(2) = 4.82, p = .09$]. Again, this result indicates that it is difficult for LSA to detect local processing strategies. Thus, the science LSA space does not generally fail to discriminate between the reading strategies, but may be limited to discriminating the extreme strategies (low vs. high integration of text with knowledge).

DISCUSSION

In this study, we evaluated LSA's ability to make automatic judgments of the quality and types of reading strategies readers use during self-explanation. Our first goal was to assess whether the domain-specific science LSA corpus can correlate with human judgments of similarity as well as or better than the LSA_GR space. This deter-

Table 2
Percentage of Correct Classifications of Test Protocols by the General Reading LSA Space and the Best Performing Science LSA Space

	Reading Style (Based on Human Codings)		
LSA Space	Minimal Processing	Local Processing	Global Processing
LSA_GR	40%	23%	59%
Science LSA 80%	38%	32%	54%

mination facilitates the decision concerning which semantic space the SERT tutor would use when making judgments of reader strategy. Also, part of our first goal was to explore which factors were key to producing a good performing corpus. Our second goal was to evaluate our science LSA spaces with respect to how they can classify reader strategies with respect to human classifications.

Our results show that the domain-specific science LSA corpus provides a good match to human ratings of the semantic relatedness of test protocols to benchmarks representing different reader strategies. LSA cosines obtained from the science spaces were highly correlated with human ratings, and more so than the LSA_GR corpus, which, in a previous analysis, had successfully matched student protocols to benchmarks that represented reading strategies (Magliano et al., 2002). The correlational analysis in the current study suggests that important factors to consider when training corpora are the amount of dimensions and formatting. Corpora trained on 350 dimensions correlated most highly with human judgments of similarity. Also, the corpora formatted into paragraphs correlated significantly better than when they were formatted into sentences. The amount of general versus specific domain knowledge in the corpus did not have an effect on the performance of the science LSA corpora. Perhaps this distinction within a domain-specific corpus does not significantly change the LSA space because both specific and general texts may contain similar information. With respect to the amount of text required to create a well-performing LSA space, it appears that the science text corpus is large enough for the present purposes, since the smaller spaces (as low as 60%) performed as well as the 100% space. This result suggests that additional text—at least from the science text domain—would not lead to noticeable improvement of the performance of this science LSA corpus.

LSA cosines were significantly different for test protocols that human experts rated as unrelated, somewhat related, or related to semantic benchmarks. This was the case for both the corpus that correlated most highly with human judgments and the corpus that correlated least highly. Thus, LSA cosines that are produced for each level of semantic similarity may be a reliable tool in discriminating between reading strategies as measured by "bags-of-words" benchmarks.

To evaluate our science LSA spaces, we examined the correspondence of their classifications of reading strategies to those of human judges, and how that correspondence compared with that of the LSA_GR corpus. Our data suggest that LSA, using both general reading and science corpora, may be useful in determining which strategy a reader is using by producing cosines between think-aloud protocols and benchmarks that represent different reader strategies. For instance, the highest cosines produced for minimal responses, as coded by human raters, were produced with the CS benchmark, and the highest cosines for global responses were produced with the WK benchmark. However, it appears as though both LSA corpora cannot distinguish local responses from minimal or global ones. This was evident by a lack of a peak in cosines for their respective benchmark (PT). It was found that the LSA_GR corpus was the only space able to produce a significant match with human judgments of processing level. One possible explanation for this finding is that the topic keywords probably influence the cosines for the science LSA because they are represented with high frequency in the LSA space. These words may, however, not reveal the difference between minimal and local processing levels. Nevertheless, the LSA_GR corpus still had difficulty classifying local processing responses. The second set of analyses were important because they highlight possible shortcomings of an LSA space trained for a specific text domain. They also qualify the advantage we found for the science LSA corpus in the correlational analyses: The science LSA may be better for some types of text analyses, whereas it looks as if a general, domain-unspecific corpus may be a better choice in the context of other tasks.

In this study, humans coded student think-aloud protocols in terms of three categories: minimal, local, and global processing. Minimal and local processing protocols were coded as the extent to which students use text information. In other words, the source from which minimal and local processing protocols may draw is the current sentence or the immediate prior text (i.e., intertextual information). The coding of global strategies often involves the extent to which students use world knowledge related to the text as well as prior text information. Thus, the LSA coding of these types of protocols involve comparing them with benchmarks that represent these three sources (CS, PT, and WK). Global processing protocols may not contain many of the keywords that are frequent in the science LSA corpus. This may explain why the science corpus does well when discriminating minimal protocols from global protocols but not local from global or minimal protocols. Possibly, local protocols tend to be confused with minimal ones because they are both based on intertextual information. However, the LSA_GR corpus may contain words indicative of global processing from outside the science domains. So, it is possible that the LSA_GR corpus is more sensitive than our science LSA corpus to the difference between local and global processing. It is important to note, however, that each space tested had the greatest difficulty in matching human coding of local processing protocols. Our data suggest that the domain-specific science space may be most useful in discriminating protocols that are based on either text information alone or world knowledge alone. It may not be able to pick up on local processing reading strategies to maintain local coherence, for example. This difficulty may mean that new techniques for representing local processing need to be explored if a program must detect these strategies. However, the results of the correlational analyses suggest that the magnitude of cosines may be a possible indicator of which strategy is most likely being employed.

In the context of the SERT tutor, these results suggest that LSA cosines may be effective in assessing the simi-

larity between student protocols and benchmarks that represent different reading strategies. This technique may allow the tutor to automatically judge which strategy a reader is using and correctly give feedback that can aid the student in using appropriate strategies for a richer understanding of the text. Finally, the comparison of the science LSA and the general reading corpus provides further support of the finding of Shapiro and McNamara (2000) that an LSA corpus trained on the specific text domain of an application provides the best match to human ratings.

REFERENCES

Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 251-282). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., de Leeuw, N., Chiu, M., & La Vancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, **18**, 439-477.

Cote, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendor & S. R. Goldman (Eds.), *The construction of mental representations during reading* (169-193). Mahwah, NJ: Erlbaum.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *Proceedings of the ED-MEDIA '99 Conference* (pp. 939-944), Seattle.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, **101**, 371-395.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, **8**, 129-147.

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

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison on latent semantic analysis and humans. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Levinstein, I., McNamara, D. S., Boonthum, C., Pillaraisetti, P., & Yadivalli, K. (in press). Web-based intervention for higher-order reading skills. In *Proceedings of the ED-MEDIA '03 Conference*, June 23-28, Honolulu.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processes during comprehension. *Journal of Educational Psychology*, **91**, 615-629.

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Muñoz, B. D., & McNamara, D. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers*, **34**, 181-188.

McNamara, D. S. (2003). *SERT: Self-explanation reading training.* Manuscript submitted for publication.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society* (pp. 387-392). Mahwah, NJ: Erlbaum.

Millis, K. K., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. (2001). Using LSA in a computer-based test of reading comprehension. In J. D. Moore, C. Luckhardt-Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-ED in the wired and wireless future*: *Vol. 68. Frontiers in artificial intelligence and applications* (pp. 583-585). Amsterdam: IOS Press.

Shapiro, A. M., & McNamara, D. S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research*, **22**, 1-36.

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes*, **21**, 255-287.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. P. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 535-542). Amsterdam: IOS Press.

(Manuscript received October 1, 2002;
revision accepted for publication March 6, 2003.)