Behavior Research Methods
2006, 38 (4), 616-627

Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts

JOSÉ A. LEÓN, RICARDO OLMOS, and INMACULADA ESCUDERO
Autonomous University of Madrid, Madrid, Spain

and

JOSÉ J. CAÑAS and LALO SALMERÓN
University of Granada, Granada, Spain

In the present study, we tested a computer-based procedure for assessing very concise summaries (50 words long) of two types of text (narrative and expository) using latent semantic analysis (LSA) in comparison with the judgments of four human experts. LSA was used to estimate semantic similarity using six different methods: four holistic (summary–text, summary–summaries, summary–expert summaries, and pregraded–ungraded summary) and two componential (summary–sentence text and summary–main sentence text). A total of 390 Spanish middle and high school students (14–16 years old) and six experts read a narrative or expository text and later summarized it. The results support the viability of developing a computerized assessment tool using human judgments and LSA, although the correlation between human judgments and LSA was higher in the narrative text than in the expository, and LSA correlated more with human content ratings than with human coherence ratings. Finally, the holistic methods were found to be more reliable than the componential methods analyzed in this study.

Discourse research has provided an increasingly precise understanding of the factors that influence the comprehension of written material, such as its structure or the role played by a reader's previous knowledge. New tools, such as latent semantic analysis (LSA), have recently been developed that could lead to an important advance in discourse research. LSA is an automatic statistical method for representing the meanings of words and test passages. A primary method for applying LSA is to use it to make predictions about the coherence of a text by comparing some units of the text (such as a sentence, paragraph, summary, or the whole text) with an adjoining unit to determine the degree to which the two are semantically related. The basic idea behind LSA is that the contexts in which words appear or do not appear provide constraints sufficient to allow one to estimate the similarities between the words. Thus LSA provides a measure of the similarities between different linguistic units. In fact, LSA permits comparison of semantic similarities between different pieces of textual information, such as sentences and paragraphs (Foltz, 1996; Landauer, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998; Landauer & Psotka, 2000) as well as summaries (Foltz, 1996; Kintsch et al., 2000).

The authors acknowledge the support of BSO 2003-06975 of the MCYT, Spain. We also thank the middle and high school students and the six experts for their participation in this study, as well as the anonymous reviewer, who provided fruitful comments. Correspondence should be addressed to J. A. León, Departamento de Psicología Básica, Universidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain (e-mail: joseantonio.leon@uam.es).

LSA is a method for extracting and representing word meanings from a large corpus of text, as well as a theory of knowledge representation (Landauer & Dumais, 1997). Although LSA as a theory of knowledge representation has been discussed by different researchers (e.g., Glenberg & Robertson, 2000; Perfetti, 1998), an emerging body of evidence supports the reliability of LSA as a tool for evaluating the semantic similarities between units of discourse; LSA has also proved comparable to human judgments of similarities in documents (Landauer & Dumais, 1997; Landauer, Foltz, et al., 1998). For example, LSA-generated cosines have been tested on a large number of essays on a diverse field of topics and have obtained a high correlation with human assessments (Landauer, Laham, & Foltz, 1998), as well as with the performance of college-bound students taking the Test of English as a Foreign Language (Landauer, Foltz, et al., 1998). LSA has also been used to determine the coherence of texts (Foltz, Kintsch, & Landauer, 1998). Last, other researchers have successfully used LSA with verbal protocols and reading strategies (Magliano, Wiemer-Hastings, Millis, Muñoz, & McNamara, 2002; Millis et al., 2004) as well as with a computerized tutoring program called AutoTutor (Graesser et al., 2000).

The Importance of a Short Summary

One area of text comprehension research that has most interested psychologists and discourse researchers concerns the processes that occur during the comprehension and summarizing phases of reading. Comprehension and

summarizing are very closely related. In fact, some researchers (e.g., Palincsar & Brown, 1984) have suggested that if readers are not able to summarize a passage, then they have not understood it. A generally acknowledged practice consists of using a summary to organize and emphasize the most relevant content of the text. Although the summary concept is imprecise, summaries themselves hold a significant place in scientific texts, and their effectiveness in improving comprehension and recall is generally recognized (Hartley & Trueman, 1983; E. Kintsch et al., 2000; León & Carretero, 1995). When readers summarize a passage, they tend to form a nucleus of information, a core concept that represents a general vision of the text in a coherent way. Synthesis and coherence are two key aspects of a good summary.

In order to summarize a text, a reader must read and comprehend the material, isolate the main ideas, and convey those ideas succinctly. In general, we can assume that a summary is a concise statement of the most important information in a text. A summary should describe most of the main ideas (or main topics) in the text. The ability to be concise is very important in some instances (e.g., when one is submitting a scientific article or a proposal for meetings or conferences, which usually require an abstract of 75 words or fewer). Because this task involves deeper processing, including writing strategies such as generalization, synthesis, and maintaining coherence (see, e.g., Brown & Day, 1983; van Dijk & Kintsch, 1983), it is more complicated than simple reading. Summarizing is especially important in educational and professional contexts (e.g., training in reading and writing strategies and assessments and in e-learning assessment, respectively).

The term *coherence* is central to discourse comprehension as well as to summarizing. Coherence is accepted as a main characteristic of a reader's mental representation of text content. Coherence relations are constructed in the reader's mind and depend on the skills and knowledge that the reader brings to the situation (Graesser, Singer, & Trabasso, 1994). A summary is considered to reflect how coherent (or incoherent) an understanding of the text the reader has. To summarize well, a reader must first perceive a text as coherent and then ensure that the ideas conveyed in the text hang together in a meaningful, organized, and synthetic manner. This analysis requires differing integrated levels of representation, including text-based models (based on topics and ideas from the text) and situational models (based on the reader's prior knowledge). As a result, summarizing is a highly effective means of constructing and integrating new knowledge. Many aspects of discourse contribute to coherence, including coreferencing, causal relations, connectives, and signals. These are highly correlated with other coherence factors such as causal relations found in the text (Fletcher, Chrysler, van den Broek, Deaton, & Bloom, 1995; Trabasso, Secco, & van den Broek, 1984).

The potential for summarization to improve comprehension is high, because it requires much more active meaning construction than choosing the best response from a set of choices or even writing short answers to isolated questions. Perhaps for this reason, as some authors suggest (e.g., Kintsch et al., 2000), summarizing may be a more authentic method for assessing what readers do and do not understand about a text than traditional comprehension tests.

A Step Forward in the Assessment of the Reliability of LSA As a Method for Grading Text Summaries

The assessment of student summaries provides a useful method for comparing LSA cosines and grades assigned by humans. For example, Kintsch et al. (2000) conducted two different comparisons of LSA scores with scores from human graders on summaries (written by 5th-grade students) of texts with an average length of 250–350 words. In the first comparison, they derived the LSA cosine between the student summaries and the text the students had read, obtaining a correlation between the teacher grade and the LSA cosine of $r = .64$. In the second comparison, they assessed whether LSA could match a given sentence from a summary to a particular section of the source text as accurately as two human graders. The LSA scores matched 84.9% of the first grader's scores and 83.2% of the second grader's. Kintsch et al. (2000) concluded that LSA scores were quite comparable to scores an experienced teacher would give to these summaries and that LSA performed almost as well as humans in determining the source of knowledge for a given sentence.

In this article, we address several issues that are important for increasing the evidence in favor of LSA as a method for assessing the quality of text summaries. We wish to explore three main questions in this study.

1. How reliable are LSA assessments when the length of the summaries is reduced to 50 words, in contrast with LSA assessments from other studies that have used longer summaries? An important question is whether the length of the summary affects its quality and, consequently, its LSA cosine. In previous studies (e.g., Kintsch et al., 2000), researchers have used summaries of 250–350 words. What would occur if we reduced the length of summary (to approximately 50 words) and asked students for more conceptualization? Would the LSA ratings still be reliable?

2. How reliable is LSA's assessment of short summaries in narrative versus expository texts? There are some reasons to think that the summarization of narrative texts differs from the summarization of expository ones, although the explanation of these differences is still under discussion. Some authors propose that readers spontaneously set in motion different patterns of activation or inference depending on the type of text they are reading (Einstein, McDaniel, Owen, & Coté, 1990; León, Escudero, & van den Broek, 2003). Narrative texts convey information about familiar events and situations in a predictable manner, whereas expository texts, by their very nature, often expose readers to new information. Differences between types of texts have also been explained in terms of the different modes of cognitive functioning they require, which correlate with two differing types of text: narrative and expository (see, e.g., Bruner, 1986; Escudero, 2004;

Goldman & Bisanz, 2002; León & Peñalba, 2002; León & Slisko, 2000; Martins, 2002; Polkinghorne, 1988).

Narrative texts make particular connections between facts, whereas an expository style tends toward the search for true, universal conditions. The narrative form usually reflects reasons, the actions of a protagonist, and the problems of daily life or fiction, and it is heavily influenced by the temporary relations that regulate the attainment of the different facts or actions. In contrast, the expository mode frequently features the conceptualization of ideas, explicitly specified rhetorical organization, context-bound terminology, and technical uses of terms. Expository discourse structure represents text types that offer conceptualizations of knowledge or ways to build knowledge (Kucan & Beck, 1996). Thus, summarizing the main ideas of an expository text becomes a different task from summarizing the plot sequence of a narrative. Synthesizing information from an expository text to construct new knowledge relations is quite distinct from summarizing a narrative text with respect to moral lessons, emotional evocation, or the actions of a protagonist. Whatever the explanation for the differences in summarization generated by text type, we think it would be interesting to examine whether LSA algorithms can detect them. Recent results presented by Wolfe (2005) suggest that LSA predicts better recall of expository texts than of narrative texts. However, it has not yet been tested whether LSA can show the same predictive differences with text summaries.

3. Finally, what is the quality of six different methods of LSA applied to two types of assessment (content and coherence), and are these methods interdependent or do they contribute independently to forming human ratings? We wanted to analyze whether all LSA methods yield equally valid results when their cosines are compared with human ratings of summaries of both narrative and expository texts that scored content separately from coherence. In particular, we analyze whether some methods are more reliable than others for content and some for coherence in narrative and expository texts. If so, how can the individual contribution from each different method be used to improve the assessment of the summaries? LSA has been tested using a variety of methods. In a series of studies, Landauer, Laham, et al. (1998) tested LSA on a large number of essays on a diverse range of topics (psychology, biology, history) for which LSA had been used to assign holistic qualitative scores. Landauer, Laham, et al. used five different methods. We compare these methods (with some minor variations on the originals) and another applied by Kintsch et al. (2000).

THE PRESENT STUDY

In this study, we are using LSA as a method for estimating the semantic similarities between sets of summaries and text, not as a theory of knowledge representation. For the objectives of this study, LSA's ability to simulate human judgments about summaries has been tested in six different ways, following the methods applied by other researchers (e.g., Foltz, Gilliam, & Kendall, 2000; Kintsch et al., 2000; Landauer, Laham, et al., 1998). These researchers have distinguished between holistic (H) and componential or analytic methods (C); all of these methods except one have been used previously to score essays. Holistic and componential methods differ in how they score the summaries. Whereas holistic methods provide a scoring of the summaries on the basis of their overall similarity to the global text (or expert summary), componential methods calculate scores on the basis of the similarity of multiple components of the summary (such as individual sentences, coherence, content, or main topics) to the global text.

According to Foltz et al. (2000), each approach has different advantages. Whereas the holistic method can typically provide a more accurate measure of the overall summary quality, the componential scoring method can provide more specific detail about which components of the summary scored better. In this study, we selected six different methods, four holistic and two componential; they are described below.

The Six Methods

Method H1: Summary–text. This holistic method consists of comparing each student's summary with all of the text that was read to derive the LSA cosine. The higher the cosine between the summary and the text is, the better the summary will score. This method has been applied by Kintsch et al. (2000) using summarization tasks in their Summary Street computerized tutoring system.

Method H2: Summary–summaries. A second holistic method consists of analyzing all of the summaries produced by students to establish similarities among all of them. Each summary is then assigned its average cosine in comparison with the average cosine for the other summaries, meaning that the summary most similar to the other summaries would receive the highest evaluation; the second most similar summary would receive the second highest evaluation, and so forth. Landauer, Laham, et al. (1998) used a similar method, but they applied the matrix of distances (1 − cosine) to student essays instead. The matrix of distances between all essays was unfolded to the single dimension that best reconstructed all of the distances, and where an essay fell along this dimension was taken as the measure of its quality.

Method H3: Summary–expert summaries. A third holistic method consists of assessing student summaries by comparing them with an expert summary. In our study, six summaries written by experts were chosen as the standard, and the LSA cosine of each student summary compared with the average LSA cosine of the six expert summaries was computed. Thus the student summary that was most similar to the expert one was evaluated as the best. A similar method was applied by Landauer, Laham, et al. (1998) to student essays.

Method H4: Pregraded summary–ungraded summary. In this final holistic method, a sample of summaries was first graded by 100 instructors, then the cosine between each pregraded summary and the remaining ungraded summaries was computed. Once the cosine was computed, each

ungraded summary was assigned the average score of a set of 10 closely similar summaries, weighted by their similarity. The main strength of this method is that it uses human judgments as the starting point. This method has been applied by Landauer, Laham, et al. (1998) to student essays.

Method C1: Summary–sentence text. This componential method consists of comparing each summary with each sentence in the text that was read. The cosine is computed by averaging the cosines between the participant's summary and all the sentences from the text.

Method C2: Summary–main sentence text. This last componential method is very similar to the previous one. It consists of computing the cosines between each sentence in a student's summary and a set of sentences from the original text that experts consider to be of importance and then averaging the cosines. This method has been applied by Landauer, Laham, et al. (1998) to student essays.

The Spanish LSA Corpus Used in This Study

The Spanish LSA database developed for this project contains documents pertaining to general topics taken from Internet resources, textbooks, online encyclopedias, newspapers, and literary books. Altogether, it contains 2,059,234 documents (i.e., paragraphs), which include 1,661,954 different terms (without syntax parsing), with the corpus finally set at 337 dimensions. We evaluated the performance of the Spanish LSA database by comparing cosines with human ratings. The Spanish LSA database is available at the University of Colorado Web site: lsa.colorado.edu.

Summary Materials and Human Expert Ratings Procedure

The summaries used for this evaluation were taken from León et al. (2004). The summaries were obtained from 390 14- to 16-year-old students attending middle or high school and six experts (PhD students). These summaries reflected content from the text and the prior knowledge of the reader, according to his or her ability and knowledge. The summaries were hand-coded by four graders who had been trained for four months. The graders scored each summary independently according to two scales: one for content (0–4 point scale) and the other for coherence (0–6 point scale), as explained below. The procedure for summaries and human expert assessments collection was as follows.

Narrative Text

Participants. Six experts (4 PhD students and 2 teachers) from the Autonomous University of Madrid and 198 middle and high school students (14–16 years old) participated voluntarily in this study.

Materials. A Spanish folktale, "La Leyenda del Algarrobo" (The Carob Tree Legend), analyzed by León and the Reading Literacy Research Group ([RLRG] 2004) in an extensive study of reading literacy, was used in this study. This narrative is 402 words long, and prior general knowledge is required to understand it.

Procedure. Each participant read the text at his or her own pace in a quiet room. The participants were required to read the text, answer two multiple choice comprehension questions, and write a concise, four-line summary with a maximum of 50 words. We chose this short summary for two principal reasons: to analyze the middle and high school students' ability to sum up the text and to analyze how well LSA cosines assess concise summaries. It should be noted that previous studies (e.g., Kintsch et al., 2000) have used summaries of 250–350 words (written by 5th grade students). In our study, the participants had a maximum of 15 min to complete the summary. They were also instructed that it was important to understand the text because they would be answering questions about it after reading it.

Human expert assessment. Four PhD students were trained for 4 months in summary assessment. This training was performed using different types of texts (narrative, expository, and argumentative), following the criteria described in León & the RLRG (2004). The evaluation of the summaries was divided into two parts. The first evaluated the content of the text on a scale of 0 to 4 on the basis of its four main components: introduction, problem, planning, and resolution. The second measured coherence on a scale of 0 (*incoherent*) to 6 (*highly coherent*). The measure of coherence involved causal relationships, topics and main idea relationships, and use of connectives.

Each grader rated each summary individually and alone, and their ratings were recorded in a statistical package (SPSS). These were then compared with the LSA cosines applying the six methods described previously.

Expository Text

Participants. One hundred ninety-two 14- to 16-year-old students from Madrid middle and high schools participated voluntarily in this study.

Materials and Procedure. "Los Árboles Estranguladores" (The Strangler Trees), an expository text analyzed by León and the RLRG (2004) in an extensive study of reading literacy, was used in this study. This expository text was extracted from a general encyclopedia that was appropriate for the general reading skills of all participants. The text contained 500 words and also required prior general knowledge. The procedure used in this study was identical to that used in the narrative study.

Human expert assessment. The four PhD students who evaluated the narrative text and summaries also evaluated the expository summaries. The criteria applied in assessing them were similar to those used for narrative texts. The evaluation was divided into two parts. The first evaluated content on a scale of 0 to 4 on the basis of its four main components: the problem of adaptation to an environment without light, the description of this type of tree, the consequences of the trees' development, and the area where they grow. The second part measured coherence on a scale of 0 (*incoherent*) to 6 (*highly coherent*). This measure of coherence involved causal relationships, topics and main idea relationships, use of connectives, and the absence of syntactic redundancy. As a preview study, each grader rated each summary individually and alone; their scores were then compared with the LSA cosines, by applying the six methods described previously.

Table 1
Correlation Matrix for LSA-Based and Human Ratings of Summaries in Narrative Text***

Method	Grader 1	Grader 2	Grader 3	Grader 4
H1. Summary–text	.55	.54	.60	.47
H2. Summary–summaries	.54	.55	.57	.49
H3. Summary–expert summaries	.52	.52	.53	.46
H4. Pregraded–ungraded summary	.57	.50	.53	.50
C1. Summary–sentence text	.58	.56	.60	.50
C2. Summary–main sentence text	.57	.55	.59	.48

$^{***}p < .001$.

LSA Assessment

As mentioned earlier, we used six different methods of LSA assessment: four holistic (H1: summary–text; H2: summary–summaries; H3: summary–expert summaries; H4: pregraded summary–ungraded summary) and two componential (C1: summary–sentence text, and C2: summary–main sentence text). The similarities among summaries, including expert summaries, were computed by measuring the cosine of the contained angle between the vectors in semantic *d*-space. The number of dimensions for this study was 337.

RESULTS

Four series of analyses were performed on the data. First, we evaluated interrater reliability among human expert assessments for each text. Second, we correlated LSA cosine scores obtained from the aforementioned six methods with human expert assessments for each type of text and each type of component assessed (content and coherence). Third, we compared those correlations in order to evaluate the relative reliability of methods for each text and each component by applying the ANOVA test (methods $\times$ text $\times$ assessment). Fourth, we performed regression analyses to evaluate the independent proportion of variance of human expert ratings explained by each method.

Interrater Reliability Test Among Human Experts' Ratings

Before we analyzed whether LSA was a reliable tool in assessing summaries, it was necessary to test the reliability among graders. For the narrative text, the correlations ranged from .79 to .84 (changed by Pearson correlation) for overall ratings. These data were used as a baseline to compare correlations between LSA cosines and graders' ratings. Interrater reliability correlations for content ranged from .81 to .86 and from .66 to .75 for coherence. For the expository text, reliability among graders ranged from .64 to .82 (Pearson) on overall ratings. Interrater reliability correlations for content ranged from .53 to .81 and from .58 to .79 for coherence.

Analysis of Correlations Between LSA Cosines and Human Experts' Ratings

In the narrative text, correlations between LSA cosines and graders' scores were obtained for each method (see Table 1). All of the correlations were positive and statistically significant ($p < .001$). For the six methods, all of the correlations between grader ratings and LSA cosines were similar; thus all methods work in a similar manner. In the narrative texts in particular, holistic methods were comparable to componential methods. The correlations found were comparable to those found by Kintsch et al. (2000) in texts pertaining to ancient civilizations.

Table 2 shows correlations between LSA cosines and the experts' scores for expository text. For the first five methods, all of the correlations are positive and statistically significant ($p < .01$). The summary–main-sentence-text method shows a nonsignificant correlation between Grader 3's ratings and the LSA cosine. With the expository text, the six methods did not work as they did with the narrative text. Some methods were more reliable than others when LSA simulated human assessment. In general, holistic methods were more reliable than componential methods in assessing the expository text in all cases studied.

The correlations between the LSA cosines and the components of assessment (content and coherence) yielded these results: In the narrative text, there were always significant positive relationships between grader ratings and the LSA cosines derived from each of the six methods.

Table 2
Correlation Matrix for LSA-Based and Human Ratings of Summaries in Expository Text

Method	Grader 1	Grader 2	Grader 3	Grader 4
H1. Summary–text	.40***	.33***	.37***	.40***
H2. Summary–summaries	.42***	.42***	.31***	.48***
H3. Summary–expert summaries	.56***	.52***	.41***	.61***
H4. Pregraded–ungraded summary	.52***	.57***	.48***	.63***
C1. Summary–sentence text	.27***	.22**	.22**	.27***
C2. Summary–main sentence text	.21**	.17*	.14	.22**

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The average correlation between human content ratings and LSA cosines was .58; between human coherence ratings and LSA cosines it was .42. Thus correlations were greater between content ratings and the LSA cosines than between coherence ratings and the LSA cosines. With the expository text, there was always a significant positive relationship between grader ratings and the LSA cosines derived from the first five methods. The sixth method showed a nonsignificant correlation between the coherence ratings given by Grader 3 and the LSA cosines. Correlations between the human ratings and the LSA cosines derived from the sixth method were lower than the other correlations. In fact, these data are also reflected in the results from the two componential methods. As previous data have shown, holistic methods were more reliable than componential methods for the expository text in all cases studied. The average correlation between human content ratings and the LSA cosine was .35; between human coherence ratings and the LSA cosine, .35. Thus correlations between the content ratings and the coherence ratings were similar. To examine the differences between texts, methods, and components of assessment, an ANOVA test was applied.

ANOVA on Correlational Data

In order to compare all of these correlational data and draw conclusions from them, we performed a 6 (methods) × 2 (type of text: expository or narrative) × 2 (type of assessment: coherence or content) ANOVA, which allowed us to answer the following questions: (1) Are the LSA assessments reliable when the length of the summaries is reduced to 50 words? (2) Is LSA a reliable tool for assessing summaries of narrative and expository texts? (3) Does the quality of the six methods differ for assessing content and coherence? The results will be described with reference to Figures 1 and 2.

Comparison of texts (narrative and expository). We found differences in the magnitudes of the correlations between the LSA cosines and human ratings for the two types of text. The correlations between human ratings and LSA cosines were found to be higher in narrative text than in expository text [$F(1,72) = 191.18$, $MS_e = 0.003$, $p < .001$]. Thus, the correlations between human ratings and LSA cosines were higher in the narrative text than in the expository text.

Comparison of the six methods. We also found differences in the magnitudes of the correlations obtained from the six methods used [$F(5,72) = 17.34$, $MS_e = 0.051$, $p < .001$]. The results show that the H2 (summary–summaries), H3 (summary–expert summaries), and H4 (pregraded–ungraded summaries) methods were the best. H1 (summary–text) was somewhat worse than H4 (pregraded–ungraded summary). The two componential methods performed the worst.

Comparison of assessment components. We also found differences in the magnitudes of the correlations between assessment components [$F(1,72) = 57.66$, $MS_e = 0.168$, $p < .001$]. These results show that LSA correlates better with human content ratings than with human coherence ratings. In other words, the similarity between human ratings and the LSA cosine is greater for content than for coherence.

Comparison of text type and the various assessment components. There was an interaction between assessment components × type of text [$F(1,72) = 51.52$, $MS_e = 0.003$, $p < .001$]. Whereas there was no difference in the average correlations between LSA cosines and human content and coherence ratings of the expository text, the difference with regard to the narrative text was statistically significant. LSA assessed the students' content summaries better than their coherence summaries in the narrative text. These data are consistent with the human ratings related to assessment of content and coherence, in which the interrater reliabilities were higher for content than for coherence in the narrative text.

Comparison of text type and method type. An interaction was found between method and text type [$F(5,72) =$

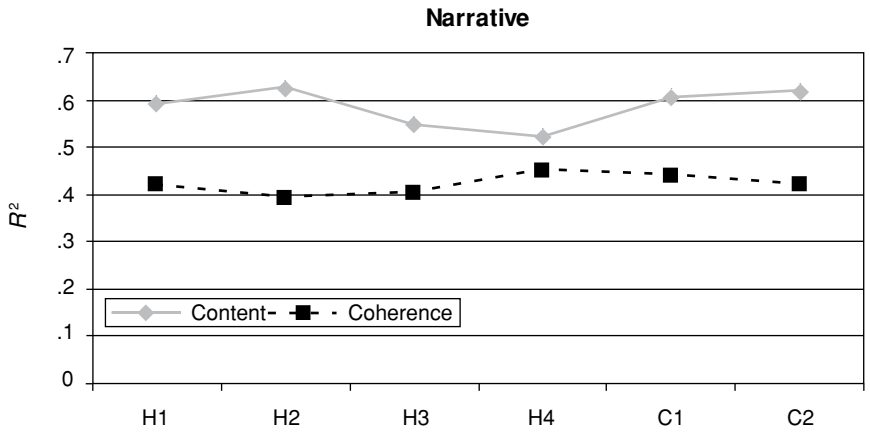

Figure 1. Interaction between content and coherence assessments and method on average correlations relative to narrative text. H1, summary–text; H2, summary–summaries; H3, summary–expert summary; H4, pregraded summary–ungraded summary; C1, summary–sentence text; C2, summary–main sentence text.

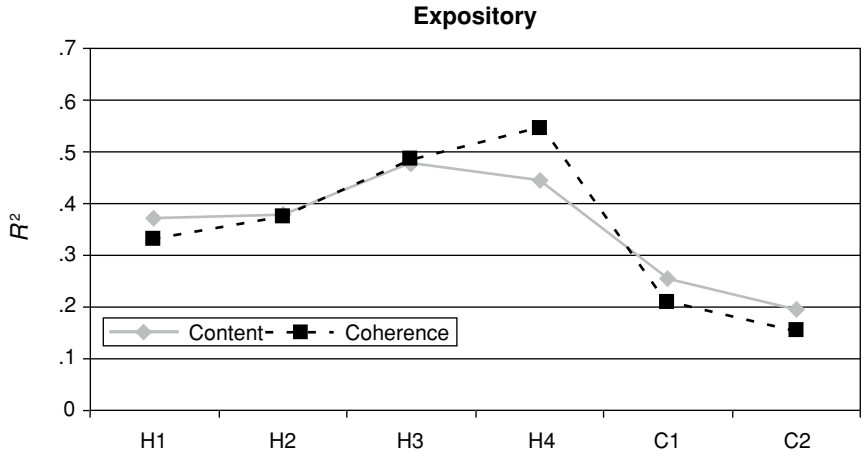

Figure 2. Interaction between content and coherence assessments and method on average correlations relative to expository text. For explanations of the abbreviations, see Figure 1.

29.34, $MS_e = 0.086, p < .001$]. For the narrative text, there were no differences among the average correlations for all six of the methods (see Figure 3); the six methods were equally reliable. However, for the expository text, there were differences among the average correlations for the six methods. We found three different means groups. A first mean related to the C1 (summary–sentence text) and C2 (summary–main sentence text) methods, which showed the lowest average correlations with human ratings, with both methods being componential. A second mean focused on the H1 (summary–text) and H2 (summary–summaries) methods, which showed higher correlations with human ratings. Finally, the H3 (summary–expert summaries) and H4 (pregraded–ungraded) methods had the highest correlations with human ratings. Contrary to the results gathered from the narrative text, where all of the different methods yielded results similar to human judgments, with the expository text reliability differed depending on the method; thus the Method section is critical when one is assessing expository texts.

Comparison between method type on each type of assessment component. An interaction was found between method and assessment components [$F(5,72) = 3.55$, $MS_e = 0.010$, $p < .05$]. The interaction was due to the statistically significant difference between average correlations for content and those for coherence found in all methods except the H4 (pregraded–ungraded summary) method.

Discussion of ANOVA on Correlational Data

In general, these results show that LSA assessment based on semantic similarity was well in line with human assessment. Judgments of semantic similarity came from comparing a summary with (1) the text, (2) summaries written by experts, (3) the remaining summaries, or, in the case of the componential methods, (4) the sentences of the text.

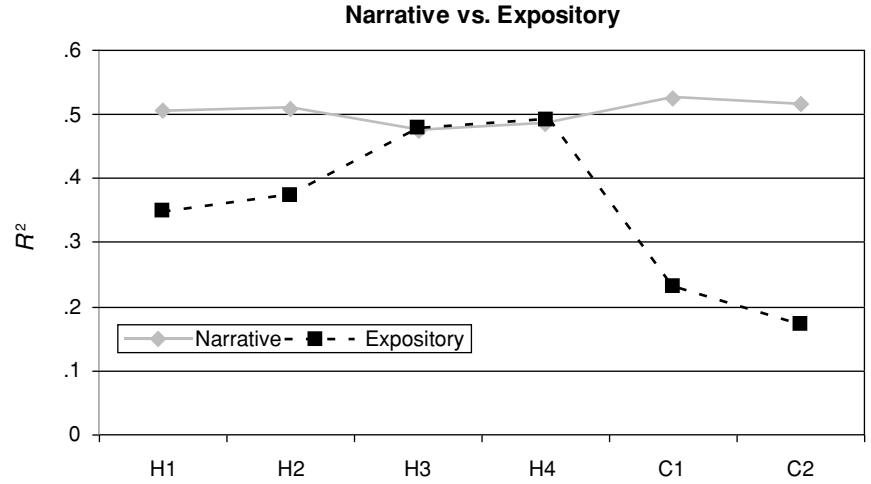

Figure 3. Interaction between type of text and method on average correlations. For explanations of the abbreviations, see Figure 1.

Therefore, semantic relationships appear to be of great importance in evaluating summary quality. These results also show that with summaries of a maximum of only 50 words, correlations between some LSA methods and human judgment are similar to correlations found in previous research (Kintsch et al., 2000). In the present study, the correlation was .64. This fact is especially reflected in the H4 (pregraded–ungraded summary) method, whose average correlation with ratings by human judges was .54. This method differs from the others in that it is based on previous assessments of some summaries done by human judges.

The H3 (summary–expert summaries) method had an average correlation of .52 with the human judges, smaller than the one obtained by Kintsch et al. (2000), in which the correlation between the teacher grade and the LSA cosine was .64. Also, it is worth noting that this method compared the student summaries with the summaries written by the six experts. Since a summary can reflect something subjective on the part of the reader, this method incorporates six expert summaries and does not make comparisons with the mental representation of only one expert.

Previous studies have shown the H3 and H4 LSA methods to be reasonably valid methods (Kintsch et al., 2000). These results are especially valid if we compare them with results obtained by Rehder, Schreiner, Wolfe, Laham, and Kintsch (1998). These authors found that the accuracy gained in the proportion of variance predicted when predicting prequestionnaire scores in essays with 200 words was five times greater than that gained in essays with 50 words. Nevertheless, the results we obtained with the H3 and H4 methods were worse (.52 average LSA correlations) than those obtained by expert human ratings (average of .78). In spite of this, the high reliability between human ratings still reflects that LSA is a reliable tool for assessing short summaries.

Moreover, taking these ANOVA results globally, we can see that correlations between LSA and human experts were higher for narrative than for expository texts. This happened mainly because four of the six methods performed worse with the expository text than with the narrative text. Only the H3 (summary–expert summaries) and H4 (pregraded–ungraded summary) methods performed equally well on both types of text. Also, it is worth mentioning that all six methods—both holistic and componential—were equally good for narrative text. Whereas for the expository text, the H3 (summary–expert summaries) and H4 (pregraded–ungraded summary) methods were significantly better than the other four methods, and the other two holistic methods, H1 (summary–text) and H2 (summary–summaries), were better than the componential methods.

The data suggest that LSA evaluated content better than coherence in the narrative text; however, in the expository text there were no differences between them. Moreover, while five methods assessed content significantly better than coherence, the H4 (pregraded–ungraded summary) method assessed both components equally well.

To answer the question about which method is best, we must consider the interaction between method and type of text and between method and type of assessment component.

Although such analyses serve to compare different methods, it is possible that all methods have some elements in common because all are similarly evaluating the same content. In order to be able to evaluate what these methods share and what they measure independently, it is necessary to perform a regression analysis to evaluate that proportion of the variance of the human expert judgments that each method explains independently. These analyses answer the second part of our third main question.

Regression Analysis

Narrative text. Eight stepwise regression models were performed on the data to evaluate how the different methods account for an independent proportion of the variance of experts' content and coherence ratings. Four stepwise regression models—one for each grader—were made to predict content, and four models were made to predict coherence. The independent variables were the six methods used in our study.

All six methods, individually considered, were statistically significant in the prediction of dependent variables. However, when we introduced some methods into the regression model, the relationship between the methods and the dependent variables disappeared because the methods did not make a unique contribution beyond what they shared with the methods already included. Table 3 shows the coefficients of the methods included in the final regression models, the statistical significance of the models, and the proportion of variance predicted for experts' content (R^2).

These results reveal some interesting aspects. In general, they show that in the narrative text, the six methods contributed more to predicting content than to predicting coherence. Proportions of the variance ranged from 38% to 46% for content and from 18% to 29% for coherence. These data are consistent with the interrater reliability of the graders' ratings (in which content ranged from .81 to .86, and coherence from .66 to .75).

A second interesting piece of data is that in all regressions analyzed, the pregraded–ungraded method showed positive and significant regression coefficients with regard to the prediction of content as well as coherence. Table 3 also shows that the summary–summaries method was included to predict three content judges' assessments. The summary–sentence-text method was included twice in the model to predict coherence. Thus, both the holistic and the componential methods are relevant and contribute to explaining the experts' assessments.

The pregraded–ungraded method appears to be the most stable and significant for both content and coherence. If we examined the data in relation to content and coherence, then the summary–summaries method, in addition to the pregraded–ungraded method, is the one that contributes most to explaining the variance of content. However, in relation to coherence, Componential Method 5 better predicts variance. Finally, in narrative texts, holistic methods explain a larger proportion of variance for coherence and content than componential methods.

Table 3
Stepwise Regression Models for LSA-Based and Human Ratings of Summaries in Narrative Text

Method	Grader 1	Grader 2	Grader 3	Grader 4
		Content		
H1. Summary–text	–	–	–	–
H2. Summary–summaries	.46***	.53***	–	.46***
H3. Summary–expert summaries	–	–	–	–
H4. Pregraded–ungraded summary	.26***	.20**	.25***	.22**
C1. Summary–sentence text	–	–	–	–
C2. Summary–main sentence text	–	–	.50***	–
F	71.05***	78.38***	83.67***	60.14***
R^2	.42	.45	.46	.38
		Coherence		
H1. Summary–text	–	–	.37***	–
H2. Summary–summaries	–	–	–	–
H3. Summary–expert summaries	–	–	–	–
H4. Pregraded–ungraded summary	.35***	.24**	.23**	.42***
C1. Summary–sentence text	.23**	.28**	–	–
C2. Summary–main sentence text	–	–	–	–
F	37.86***	27.73***	39.98**	42.82***
R^2	.28	.22	.29	.18

**$p < .01$. ***$p < .001$.

Expository text. We performed eight stepwise regression models to evaluate how the different methods account for an independent proportion of the variance of experts' content and coherence ratings. Four were carried out to predict content, and four were performed to predict coherence. The independent variables were the six methods used in this study (see Table 4).

With the expository text, the proportion of variance that accounted for content ranged from 22% to 35%; for coherence, from 23% to 50%. These results show that in the expository text, the six methods better predict coherence than content. In all of the regression analyses, the pregraded–ungraded method showed positive and significant regression coefficients with regard to the prediction of the content and coherence, as with the narrative text.

The results were similar to those found with the narrative text. However, contrary to what occurred with the narrative text, the regression weight of the H4 method reflected greater importance in predicting the dependent variable. The summary–expert summaries method was included to predict content and coherence, with grader assessments being the second most important method of the regression model. Another difference with respect to the analysis of narrative text is that all of the holistic methods were included in at least one regression model, but componential methods were not. Thus componential methods

Table 4
Stepwise Regression Models for LSA-Based and Human Ratings of Summaries in Expository Text

Method	Grader 1	Grader 2	Grader 3	Grader 4
		Content		
H1. Summary–text	–	–	.24***	–
H2. Summary–summaries	–	–	–	.38***
H3. Summary–expert summaries	.37***	.33***	–	–
H4. Pregraded–ungraded summary	.24**	.28***	.33***	.37**
C1. Summary–sentence text	–	–	–	–
C2. Summary–main sentence text	–	–	–	–
F	38.61***	37.93***	26.06***	51.11***
R^2	.29	.29	.22	.35
		Coherence		
H1. Summary–text	–	–	.19**	–
H2. Summary–summaries	–	.26***	–	–
H3. Summary–expert summaries	.36***	–	–	.33***
H4. Pregraded–ungraded summary	.32***	.51***	.38***	.47***
C1. Summary–sentence text	–	–	–	–
C2. Summary–main sentence text	–	–	–	–
F	52.57***	59.90***	27.64***	91.74***
R^2	.36	.39	.23	.50

**$p < .01$. ***$p < .001$.

did not contribute to predicting the dependent variable in the expository text. The differences found between the narrative and expository texts are consistent with the results found in the ANOVA. Another significant fact is that the H3 method appears together with the H4 method in four of the eight regression models, becoming the second most important method for predicting content and coherence. Methods 1 and 2 appear only twice in the regression models.

Therefore, the results found with the expository text are consistent with those found with the ANOVAs. There, the H3 and H4 methods most closely resembled human opinions, and in the regression these methods contributed the most weight to predicting human trials.

The results of the regression analysis highlight the versatility of LSA in assessing summaries. For example, the analysis shows that some methods, such as the pregraded–ungraded method, stay stable and consistent independently of the texts, the judgment of the different human experts, and the type evaluation (content or coherence). Furthermore, LSA is also sensitive to each of the studied variables. With regard to holistic and componential methods, the results show that the former predict expositive texts better, whereas with the narrative text, the componential methods are also important. In relation to the judges' assessments, we found that with respect to one particular judge, LSA performed better with one specific method (Judge 3 with Method H1). With regard to content and coherence, Method H2 contributed with a significant weighting to predicting the human judgments only in content with the narrative text, but with the expositive text, Method H3 worked well for both content and coherence.

Discussion of the Regression Analysis Results

The regression analyses examined to what extent the different methods of LSA are supplemented. The ANOVAs analyzed the methods by themselves, revealing the individual reliability of each one. The question that arises is whether combining the methods can contribute to predicting the human judgments even better. If the methods were mutually exclusive, a stepwise regression analysis would choose only the best method, ignoring the remaining ones. On the other hand, as long as the methods are complementary, the proportion of variance predicted by each one would be detected. The results showed a consistent pattern in the two text types.

The pregraded–ungraded summary method was always part of the regression models. Also, in all of the 16 models except one, it was accompanied by another method in the final equation of the model. Therefore, it seems that the pregraded–ungraded summary method is supplemented with the other methods and contributes something more when predicting human judgments. The remaining five methods overlapped with each other, so they competed in the regression equation until only one entered into it. It is possible that the differential contribution of the pregraded–ungraded summary method is in fact a human preevaluation made with some summaries and that this is what the method is reflecting. The other five methods, as long as they depend solely on LSA without being mediated by human judgment, explain similar percentages of variance. Furthermore, it is interesting that these five methods behave differently with narrative and expository texts. Both the componential and the holistic methods contributed individually to predicting human judgments in the narrative text. However, in the expository text, none of the componential methods entered into the equation, and the summary–expert summaries method appeared in the equation only 50% of the time. We therefore conclude that using the pregraded–ungraded summary method accompanied by a second method best predicts human judgments.

In general, the average proportion of variance correctly predicted in both texts was 34%, ranging in some cases up to 50% (the equivalent to the variance shared by two human judges who correlate .70 in their judgments). Nevertheless, the regression models predicted content better in the narrative text and coherence better in the expository text. It is worth noting that the ANOVA found that the methods corresponded more with the human judgment of narrative text. However, the regression analysis showed that when the two more successful methods were combined, the percentage of explained variance of expository texts reached that of narrative text. In other words, combining several methods improved the prediction of human judgments more in the expository than in the narrative text.

GENERAL CONCLUSIONS

Our purpose in the present study was to address several questions regarding the reliability of LSA as a computer-based procedure for assessing 50-word summaries. We gathered data from four expert human judges, two types of text (narrative and expository), and six different methods of assessment. We expected to find evidence supporting the reliability of LSA as a tool for evaluating the semantic relatedness of units of discourse in contexts and with materials on which it had not previously been tested.

An important result of our study, and one that is related to our first objective, concerns whether the length of a summary is determinant in assessing its quality in relation to its LSA cosine. In this aspect, our correlations were similar to those found by Kintsch et al. (2000) in their study of narrative texts relating to ancient civilizations. There are two main differences between Kintsch's study and ours. The first is that the summaries we used were 50 words long, as opposed to the 250- to 350-word summaries that Kintsch et al. used. It is well known that LSA does not work well when the number of words is fewer than 200 (Rehder et al., 1998). The second difference is the academic level of the participants. Our study used students from middle and high school (14–16 years old), whereas Kintsch et al. used 5th-grade students (10–11 years old). An interpretation of our results would be that restrictions on text length are compensated for by a greater conceptualization of the summary and by more concentration of key information or main topics contained in the texts. This viewpoint supports the idea that LSA is sensitive to semantic information in terms of conceptualization and abstraction.

Our second objective was to test whether summarization of narrative texts differed from summarization of expository texts, and whether the possible differences could be detected by LSA. The results show that there were differences in the way the methods behaved with respect to narrative and expository texts. Thus, correlations between human ratings and LSA cosines were higher in the narrative text than in the expository text, with the correlations greater for content than for coherence. To our knowledge, only one previous study has compared LSA performance in narrative and expository texts. Wolfe (2005) found that in recall tasks, LSA performed better with expository than with narrative texts, but our results showed the opposite pattern. However, as Wolfe (2005) suggested, the genre of a text triggers processing strategies that vary depending on the associations that are relevant to the task that people have to carry out. Already, some studies have tested how well LSA can predict recall performance. For example, Steyvers, Shiffrin, and Nelson (2004) found that LSA had problems predicting recall data. These researchers also suggested that in recall tasks, there are some retrieval mechanisms that work in ways that LSA associations cannot explain. Therefore, a subject for further research would be to examine these differences in LSA performance in differing tasks with narrative and expository texts. Furthermore, it is important to note that we used the maximum number of dimensions available for the Spanish language in the LSA space web, with the corpus finally set at 337 dimensions for narrative and expository summaries. Another question for further research would be to analyze the appropriate number of dimensions that should be used for the best quality assessment in the narrative and expository summaries.

Finally, we come to the question of the relative reliability of the different LSA-based methods for calculating summary quality. First, the comparison among these methods showed that they all performed similarly well with narrative texts, with correlations similar to those found by Kintsch et al. (2000). However, with expository texts, the componential methods clearly performed worse. We can also express this by saying that the methods that used text information to evaluate the summaries performed worse. The pregraded–ungraded summary, the summary–expert summaries, and the summary–summaries methods were clearly better. These three methods used only information contained in the summaries to make their evaluations. The three worst methods used information based on the text. Furthermore, LSA correlates better with the judges' assessments in the evaluation of content than coherence with the narrative text. However, with the expository text we found the opposite results. As Figure 1 shows, these differences could be due to how LSA evaluates content in the narrative text. The evaluations of coherence and content of the expository text and the evaluation of coherence of the narrative text are practically the same. In fact, differences between coherence and content were not statistically significant.

If these data show that holistic methods were more reliable than componential methods, they also support the idea that LSA can provide a more accurate measure of the overall quality of a summary as opposed to the quality of individual components of a summary. This viewpoint also suggests that LSA is more sensitive to evaluating how semantic information is processed in terms of conceptualization and abstraction than the componential methods analyzed.

A question that could be asked is whether there is any commonality among these LSA-based methods in explaining the variance in summary ratings. This question is relevant when one attempts to define what these methods have in common and what is different in each of them in relation to what they are measuring. The results show that the pregraded–ungraded summary method, accompanied by a second method, could be used to predict a large proportion of variance in human judgments.

In overall terms, these data support the reliability of LSA as a tool for comparing semantic similarity and human judgment in summarization. Furthermore, LSA is able to make accurate evaluations of summaries even when the summaries are no longer than 50 words. Such data also suggest that LSA is more than merely a good semantic tool. LSA has obtained successful results in differing types of text. This could mean that LSA is able to detect how human experts assess the quality of summaries. In other words, LSA seems to have predictive power. However, more research is needed in order to find out how our results could be used to test LSA as a psychological theory of text comprehension. Our results are a good starting point for studying how semantic information is processed in terms of conceptualization and abstraction, rather than in terms of syntactic or grammatical structure. This is the goal of the research we are conducting on the basis of these data.

REFERENCES

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning & Verbal Behavior*, **22**, 1-14.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Coté, N. C. (1990). Encoding and recall of texts: The importance of material appropriate processing. *Journal of Memory & Language*, **29**, 566-581.

Escudero, I. (2004). *Procesamiento de inferencias elaborativas en la comprensión del discurso y según el tipo de texto* [Elaborative inference processing in discourse comprehension and type of text]. Unpublished doctoral dissertation, Universidad Autónoma de Madrid. Fletcher, C. R., Chrysler, S. T., van den Broek, P., Deaton, J. A., & Bloom, C. P. (1995). The role of co-occurrence, coreference, and causality in coherence of conjoined sentences. In R. F. Lorch, Jr. & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 203-218). Hillsdale, NJ: Erlbaum.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, **28**, 197-202.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, **8**, 111-127.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, **25**, 285-307.

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

GOLDMAN, S., & BISANZ, G. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 19-50). Mahwah, NJ: Erlbaum.

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.

HARTLEY, J., & TRUEMAN, M. (1983). The effects of headings in text on recall, search and retrieval. *British Journal of Educational Psychology*, **53**, 205-214.

KINTSCH, E., STEINHART, D., STAHL, G., MATTHEWS, C., LAMB, R., & THE LSA RESEARCH GROUP (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, **8**, 87-109.

KUCAN, L., & BECK, I. L. (1996). Four fourth graders thinking aloud: An investigation of genre effects. *Journal of Literacy Research*, **28**, 259-287.

LANDAUER, T. K. (1998). Learning and representing verbal meaning: The latent semantic analysis theory. *Current Directions in Psychological Science*, **7**, 161-164.

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., FOLTZ, P. W., & LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

LANDAUER, T. K., LAHAM, D., & FOLTZ, P. W. (1998). *Computer-based grading of the conceptual content of essays*. Unpublished manuscript.

LANDAUER, T. K., & PSOTKA, J. (2000). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments*, **8**, 73-86.

LEÓN, J. A., & CARRETERO, M. (1995). Intervention in comprehension and memory strategies: Knowledge and use of text structure. *Learning & Instruction*, **5**, 203-220.

LEÓN, J. A., ESCUDERO, I., & VAN DEN BROEK, P. (2003). La influencia del género del texto en el establecimiento de inferencias elaborativas [The influence of type of text on the establishment of elaborative inferences]. In J. A. León (Ed.), *Conocimiento y discurso: Claves para inferir y comprender* (pp. 153-170). Madrid: Pirámide.

LEÓN, J. A., & PEÑALBA, G. E. (2002). Understanding causality and temporal sequence in scientific discourse. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 155-178). Mahwah, NJ: Erlbaum.

LEÓN, J. A., & THE READING LITERACY RESEARCH GROUP (2004). *La competencia lectora y los procesos de comprensión: Un proyecto de investigación basado en la evaluación de los tipos de comprensión* [Reading literacy and reading processes: A research project on assessment of types of comprehension]. Unpublished manuscript.

LEÓN, J. A., & SLISKO, J. (2000). La dificultad comprensiva de los textos de ciencias: Nuevas alternativas para un viejo problema educativo [The difficulty of understanding science texts: New alternatives for an old education problem]. *Psicología Educativa*, **6**, 7-26.

MAGLIANO, J. P., WIEMER-HASTINGS, K., MILLIS, K. K., MUÑOZ, B. D., & MCNAMARA, D. [S.] (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers*, **34**, 181-188.

MARTINS, I. (2002). Visual imagery in school science texts. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 73-90). Mahwah, NJ: Erlbaum.

MILLIS, K. K., KIM, H.-J. J., TODARO, S., MAGLIANO, J. P., WIEMER-HASTINGS, K., & MCNAMARA, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, **36**, 213-221.

PALINCSAR, A. S., & BROWN, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition & Instruction*, **1**, 117-175.

PERFETTI, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, **25**, 363-377.

POLKINGHORNE, D. (1988). *Narrative knowing and the human sciences*. Albany: State University of New York Press.

REHDER, B., SCHREINER, M. E., WOLFE, M. B. W., LAHAM, D., LANDAUER, T. K., & KINTSCH, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, **25**, 337-354.

STEYVERS, M., SHIFFRIN, R. M., & NELSON, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237-249). Washington, DC: American Psychological Association.

TRABASSO, T., SECCO, T., & VAN DEN BROEK, P. (1984). Causal cohesion and story coherence. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 83-107). Hillsdale, NJ: Erlbaum.

VAN DIJK, T. A., & KINTSCH, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

WOLFE, M. B. W. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 359-364.

(Manuscript received May 18, 2005;
revision accepted for publication August 19, 2005.)