Behavior Research Methods
2009, 41 (3), 944-950
doi:10.3758/BRM.41.3.944

New algorithms assessing short summaries in expository texts using latent semantic analysis

Ricardo Olmos, José A. León, Guillermo Jorge-Botana, and Inmaculada Escudero
Universidad Autónoma de Madrid, Madrid, Spain

In this study, we compared four expert graders with latent semantic analysis (LSA) to assess short summaries of an expository text. As is well known, there are technical difficulties for LSA to establish a good semantic representation when analyzing short texts. In order to improve the reliability of LSA relative to human graders, we analyzed three new algorithms by two holistic methods used in previous research (León, Olmos, Escudero, Cañas, & Salmerón, 2006). The three new algorithms were (1) the *semantic common network algorithm*, an adaptation of an algorithm proposed by W. Kintsch (2001, 2002) with respect to LSA as a dynamic model of semantic representation; (2) a *best-dimension reduction measure* of the latent semantic space, selecting those dimensions that best contribute to improving the LSA assessment of summaries (Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007); and (3) the *Euclidean distance measure*, used by Rehder et al. (1998), which incorporates at the same time vector length and the cosine measures. A total of 192 Spanish middle-grade students and 6 experts took part in this study. They read an expository text and produced a short summary. Results showed significantly higher reliability of LSA as a computerized assessment tool for expository text when it used a *best-dimension* algorithm rather than a standard LSA algorithm. The *semantic common network* algorithm also showed promising results.

Latent semantic analysis (LSA) is a computational linguistic model that offers a mathematical representation of a semantic domain. It can be also conceived of as an automatic statistical method for representing the meaning of words and text passages. This tool is capable of analyzing a huge dimensional matrix where each row represents a digitalized word (term) and the column has one paragraph (document). After that, LSA reduces the original matrix via SVD, a mathematical technique that reduces the dimensionality of a matrix, in a new semantic space where each word and each document are represented as a single vector. It has been widely shown that this reduced semantic space preserves the semantic relations between words and documents, as humans do. In this semantic space it is possible to compare units of a piece of information (sentence, paragraph, summary, or whole text) with adjoining units of the text to determine the degree to which both are semantically related. In fact, LSA permits comparison of semantic similarity between different pieces of textual information, such as sentences or paragraphs (Foltz, 1996; Landauer, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), as well as summaries (Foltz, 1996; E. Kintsch, Steinhart, Stahl, & LSA Research Group, 2000; León, Olmos, Escudero, Cañas, & Salmerón, 2006). LSA measures the similarity between two pieces of text with the cosine between the two vectors. Thus, if the cosine is near 1, the two pieces of text are very semantically similar, and if the cosine is near 0, the two pieces are not semantically related at all. However, most of the applications with LSA have conceived of this tool from a static point of view and do not take advantage of all available mathematical information that LSA contains. The present study is theoretically motivated by authors such as W. Kintsch (2001, 2002) or Denhière, Lemaire, Bellissens, and Jhean-Larose (2007). These authors have proposed extending LSA as a model not only of acquisition and semantic representation, but of semantic memory and language processing also; these are closer to human cognitive dynamic processing of texts than is the standard LSA, which gives a static representation of the semantics of the text. We are also interested in some mathematical extensions using LSA on the basis of the ideas of Hu, Cai, Wiemer-Hastings, Graesser, and McNamara (2007) that we have applied in assessing short summaries with LSA. The final objective is to obtain new algorithms that improve the quality of LSA assessment over traditional LSA methods. We intend to apply the new ideas to extend LSA psychologically or mathematically for assessment purposes, and we hope that LSA users can take advantage of the new algorithms or try to implement other models based on similar principles.

During the last two decades, research in discourse processes has focused on factors that influence language comprehension, such as the role of readers' previous knowledge, type, nature, and structure of written discourse (narrative, expository, argumentative). However, there is a lack of research on computer tools capable of accurately

J. A. León, joseantonio.leon@uam.es

assessing written discourse. New tools such as LSA could represent an important advance in discourse assessment research. One area of text comprehension research that has most interested discourse researchers concerns the processes that occur during the comprehension and summary phases of reading. For example, 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. The potential for summarization to improve comprehension is high, because summarizing requires much more active meaning construction than choosing the correct answer in a test, or even writing short answers to isolated questions. Perhaps for this reason, as some authors suggest (e.g., E. Kintsch et al., 2000), summarizing may be a more authentic method than traditional comprehension tests of assessing what readers do and do not understand about a text.

Recently, some authors have pursued two main objectives related to summaries and LSA: to provide (1) an automatic computerized tool that teaches students to write a summary (e.g., E. Kintsch et al., 2000); and (2) an automatic tool capable of assessing the quality of a summary in a manner comparable to that of human graders (Foltz, 1996; Landauer, 1998; Landauer & Dumais, 1997, Landauer, Foltz, & Laham, 1998; León et al., 2006). We have studied the second aspect; admittedly, our summaries were very short (between 25 and 50 words). Rehder et al. (1998) found that the accuracy gained in the proportion of variance accounted for when predicting prequestionnaire scores in essays of 200 words was five times greater than in compositions of 50 words. Wiemer-Hastings, Wiemer-Hastings, and Graesser (1999) showed that LSA does better with more than 60 words, and that it encounters particular difficulties in the 2- to 60-word range. In our previous study (León et al., 2006), we found an acceptable degree of relative reliability between LSA and human graders, but still far from reliability within graders. In this study, we used an expository text that middle-grade students summarized. The summaries were only moderately well assessed, so we chose the expository text as a standard by which to compare the new algorithms proposed.

In view of these results, some improvements are needed to obtain higher reliabilities in LSA assessments. In the last few years, interesting progress in LSA has been made by authors trying to give LSA more psychological plausibility (Denhière et al., 2007; W. Kintsch, 2001, 2002) as well as by authors proposing to extract more mathematical information from the semantic space (Hu et al., 2007). In the first group, two excellent examples are in Denhière et al.'s study, which conceives of LSA as a semantic space that models children's semantic memory; and in W. Kintsch's (2001) study describing the prediction algorithm, which makes language more context dependent—improving, among other things, the polysemy problem inherent in LSA (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). This implies that LSA can be seen as a dynamic model of the semantic representation getting better results. In the second group, some authors have proposed new mathematical extensions (Hu et al., 2007; McNamara, Boonthum, Levinstein, & Millis, 2007). Hu et al.'s study presents the "adaptive method" to make a more efficient use of the latent semantic space when LSA is assessing physics protocols from students. Another example is presented in McNamara et al.'s study using a weighting algorithm to reflect the different importance that the words have in different sentences. With this contribution in mind, we propose in this study to introduce new algorithms that reflect the mathematical–psychological models proposed by these authors in the context of summary assessment. These algorithms are described below.

The Study and Objectives

In this article, we tested a computer-based procedure for assessing short summaries using LSA combined with four expert human judgments in an expository text. The objective was to analyze how new algorithms could improve the reliability of LSA with human graders when assessing short summaries, compared with standard LSA use in expository text. This study is an extension of León et al. (2006), in which LSA was used with six standard methods (four holistic and two componential) in order to compare the assessment of very short summaries of narrative and expository texts by LSA with those by expert graders. The results supported a good reliability of LSA in narrative text, but only moderate with expository texts. These results raise two questions, the first about the type of cognitive demands that each type of text requires to make a summary, and the second about the methods used in the evaluation of the summaries by LSA. Concerning the role of text type, a possible explanation of these results is related to the idea that summarizing the main ideas of an expository text is a different task from summarizing the plot sequence of a narrative. Synthesizing information from an expository text to construct new knowledge relations is quite different from summarizing a narrative text with respect to moral lessons, emotional evocation, or the actions of a protagonist. Regarding the methods applied in LSA, another possible explanation is that, in general, holistic methods were more reliable than were componential methods, and their behavior would be different in narrative or expository texts. We will analyze this question in this study. Whereas holistic methods provide a unique measure of the overall quality of a summary, componential methods give several quality measures based on multiple semantic features in the summary. In León et al., we concluded that LSA was more sensitive than the componential methods analyzed to evaluating how semantic information is processed in terms of conceptualization and abstraction. These data also showed that two holistic methods (the summary–text method and the summary–expert summaries method) obtained moderately good results in expository text. Both methods were selected for the present study because they capture different semantic similarities when they compare each student summary with the whole text (summary–text method), or each student summary with the expert summary aver-

age (summary–expert summaries method). Moreover, the reason to choose the summary–expert summaries method is that it showed the best results in our previous study (León et al., 2006) as a method depending only on LSA and does not need any prior information from human graders (unlike the pregraded–ungraded method, which also works especially well, but which with LSA requires a pool of summaries previously scored by the human graders). We chose the summary–text method because it is the simplest method, and it has worked reasonably well in previous research (E. Kintsch et al., 2000; León et al., 2006). Thus, this last method is especially useful for any LSA user. We now describe briefly the two holistic methods.

The summary–text method is based on the idea that the cosine measure can capture the semantic similarity between the student summary and the text. It consists of comparing each student summary with the whole text that was read to derive the LSA cosine. The higher the cosine between the summary and the text, the higher the summary will score. This method has been used successfully by E. Kintsch in the summary street tool (E. Kintsch et al., 2000). Summary–expert summaries consist of assessing student summaries by comparing them with an expert summary (Landauer, Laham, & Foltz, 1998). It is conceived as a method that can capture how similar semantically a student summary is to other summaries written by experts, usually known as *golden summaries*. For the present study, six summaries written by experts were chosen as the standard. With this method, LSA gives a score to one student summary as follows: LSA computes cosines between the student summary and each of the six expert summaries, so six cosines are first computed. The final score to the student summary is the average of the six cosines.

Furthermore, the two holistic methods selected (summary–text and summary–expert summaries) were combined with three new algorithms to improve the reliability of LSA and human graders in expository texts, as well as to capture some mathematical and psychological extensions of LSA that make it a more useful tool. These algorithms are the following.

The semantic common network algorithm. LSA is not capable of distinguishing multiple senses of a word, since a single vector represents only one word. This is known as LSA's *polysemy problem* (Deerwester et al., 1990). W. Kintsch (2001) showed that the LSA model could be used to provide a good semantic representation (the algorithm is applied to sentences with the structure *argument–predicate*), as long as the specific role of the predicate is taken into account. The essence of the algorithm is strengthening features of the predicate appropriate for the argument. In other words, this algorithm extracts a context-dependent meaning; for example, in LSA's representation of the sentence "this lawyer is a shark," W. Kintsch (2000, 2001) proposed that only neighbors of the predicate associated with the context need to be considered. Therefore, associated neighbors such as *aggressive*, *predatory*, or *tenacious* would be activated, but not *fish*, *swimmer*, or *gills*, because, although they are close neighbors of literal shark-properties, they are not related to the argument. In this way, the algorithm incorporates information about neighbors, so that the information in *shark* is linked in a semantic network to *lawyer*.

We used the same idea to establish a common semantic network between the summary by a student and the summarized text. The general idea of our adaptation of this algorithm was to provide additional semantic information in the summary vector. Thus, if we had a summary, instead of representing the vector with the sum of its words we added to the summary its closest neighbors, expanding the semantic network. Now, the summary was composed of its own words and others semantically related to it. Psychologically, the algorithm means that when we express something in a piece of language, the meaning conveyed is more than is expressed explicitly. Therefore, this algorithm in our study means that the final vector represented in the LSA space consists of the words of the summary and the most semantically related concepts. We adapted this algorithm because instead of adding to the student summary its n closest neighbors, we only added to each student's summary the p terms most related to the expository text (where $p < n$). In particular, the semantic common network algorithm first extracted the 50 nearest neighbors of the summary. Then, we included in the semantic network not the 50 (n) neighbors but the 20 (p) most closely related to the expository text, where $p < n$, following W. Kintsch's (2007) criteria. Then we chose $n = 50$ and $p = 20$; thus, 20 terms were added to the vector summary, and these terms were semantically related both to the summary and the expository text. Thus, we spread out the semantic network by (1) activating the closest neighbors of the summary; (2) suppressing neighbors not related to the text; and (3) retaining those neighbors with relatively strong links with the text.

The best-dimension algorithm. Instead of using all the semantic space to represent a text, Hu et al. (2007) used only dimensions that best contribute to improving the LSA assessment's verbal protocols. This algorithm supposes an intelligent and discriminative use of the semantic space. In our case, we applied it to summaries by randomly selecting 30 out of 192 summaries. We chose only 30 summaries randomly to train this method, and the remaining 162 to validate it, in order to avoid overfitting the algorithm to the whole sample. The four graders rated them and we noted the average grade in each summary; we thus obtained 30 ratings previously assessed by graders. LSA rated these 30 summaries as follows: First, it removed the dimension that made the worst Pearson correlation between LSA–average human grader, and obtained the best $p - 1$ dimension semantic space in terms of LSA–human graders' reliability. Second, it removed the worst dimension in this reduced space that made the LSA–human grader reliability poorer. The algorithm continued until 20% of the worst dimensions were removed. The algorithm gave us the semantic space that most contributed to the LSA–human graders' reliability. Thus, each summary vector had 80% of the original information and LSA used

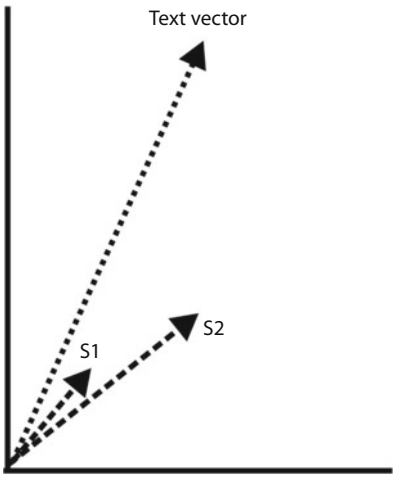

Figure 1. An example of a cosine measure overestimating a rating.

only the most relevant features of the semantic space, in much the same way as human graders consider only the most relevant features when assessing summaries.

The Euclidean distance. The cosine has usually been the measure used by LSA to evaluate texts (more generally, to evaluate similarity between texts). As a consequence, we saw that some summaries, when insufficiently elaborated, were considered too similar to expert summaries or to the summarized text. In those cases, the summary ratings provided by LSA were overestimated. For example, suppose that we use the summary–text method to assess summaries where we compare the student summary with the text summarized. Figure 1 represents this case graphically—the big arrow represents the vector of the text, and the small arrows represent two student summaries, called S1 and S2. Note that S1 is closer to the text vector than S2 is, in terms of the cosine (less of an angle), but if we consider the Euclidean distance (the distance between the end of the arrows), S2 is closer to the text than S1 is.

As an alternative solution to the cosine problem, we used the Euclidean distance measure (see a description of this and other measures in Rehder et al., 1998). This measure incorporates both vector length and the cosine. Vector length contains the quantity of summary elaboration, and the cosine contains the quantity of semantic similarity. Thus, Euclidean distance is an algorithm that contains more information about the summary contents, and it probably improves the reliability of LSA for assessing summaries. We think this measure can be especially sensitive to the method used to assess summaries. Euclidean distance would give different quality when we compare a student summary with the text (summary–text method) and when we compare a student summary with an expert summary (summary–expert summaries method). Since an expert summary has approximately the same level of elaboration as a student summary (same vector length), Euclidean distance would not be an appropriate and sensitive measure with the latter method.

METHOD

The Spanish LSA database developed for this study contains 372 documents with similar contents to the expository text used in the study. These documents were taken from Internet resources, textbooks, and online encyclopedias. This includes 5,995 lemmatized words. The semantic space was set at 75 dimensions, which cover 40% of the total variance (this criterion was proposed by Wild, Stahl, Stermsek, & Neumann, 2005).

The summaries used for this evaluation were taken from León and the Reading Literacy Research Group (2004). The summarized expository text was "Los Árboles Estranguladores" ("The Strangler Trees"). This expository text was taken from a general encyclopedia adapted to the general reading skill of all participants. It contained 500 words and also required prior general knowledge. The summaries were obtained from 192 students (age range, 14–16) attending middle/high school and 6 experts (PhD students). The summaries had a maximum length of 50 words. The 192 summaries were rated by four expert graders on a 0–10 point scale. The LSA rating, as we explained earlier, was conducted with three new algorithms (common semantic network, best dimensions, and Euclidean distance) and with a standard use of LSA. The standard use of LSA was taken as the baseline and we used it to compare the other algorithms. All three algorithms derive distinct vectors for each summary and the standard has the usual vector for each summary. To rate each summary, we compared each vector with the text vector or with six expert summaries (the six comparisons were averaged into one score).

To perform the data analysis, we applied a two-way ANOVA where the dependent variable was the correlation (Pearson) between LSA cosine and human experts. The first factor was algorithm (three new algorithms and the standard or baseline): Kintsch adapted algorithm; best-dimension algorithm; Euclidean distance algorithm; and standard algorithm, called *baseline*. The second factor was method (the two holistic methods described above): summary–text and summary–expert summaries.

RESULTS

We did not find an effect of method [$F(3,24) = 0.11$, $MS_e = 0.02$, $p = .74$], but we found differences in the magnitude of reliabilities, depending on the algorithm [$F(3,24) = 11.50$, $MS_e = 0.02$, $p < .05$]. However, these results are modulated by the interaction effect (see Figure 2).

There was an interaction effect between algorithm and method [$F(3,24) = 7.21$, $MS_e = 0.02$, $p < .05$]. The interaction was caused by the Euclidean algorithm: Failing to increase the reliability mean in the expert summaries method, this algorithm does not seem to work well. The simple interaction effects were as follows: We found that the best-dimension algorithm had more reliability than the baseline in the summary–text method and in the summary–expert summaries method ($p < .05$). We did not find any other means differences among algorithms within the summary–text method ($p > .05$). The semantic common network algorithm and the baseline algorithm did not differ significantly, in spite of the fact that the semantic common network algorithm has higher reliability means in both methods. Probably, there are no significant differences due to the lack of power of this statistical test (four cases per group). Finally, Euclidean distance reliability only behaved the same way as the rest of the algorithms in the summary–text method ($p > .05$), but gave its

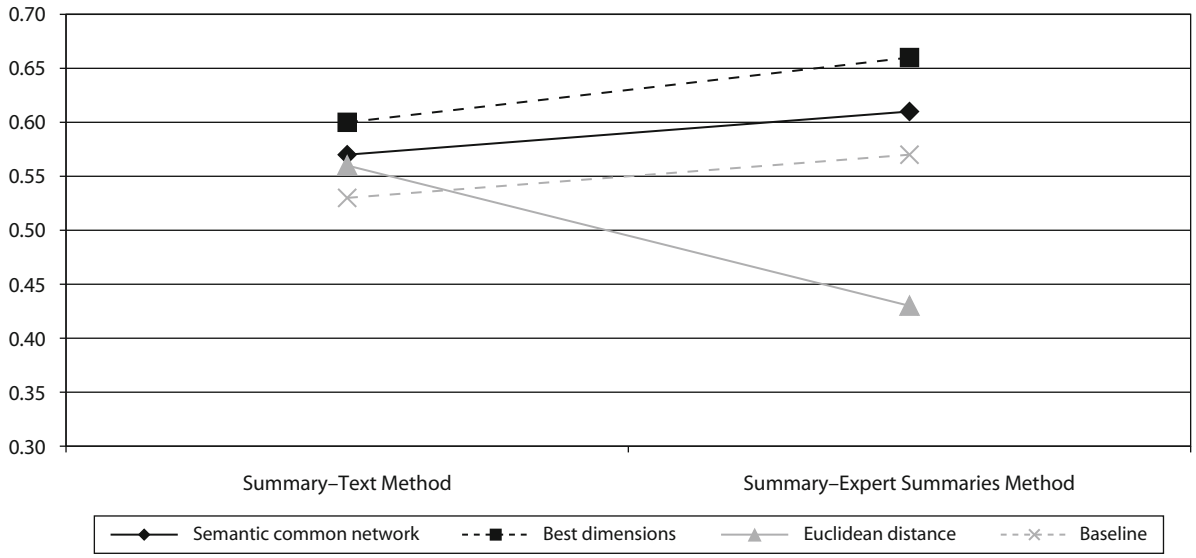

Figure 2. Interaction effect between algorithm (lines) and method (horizontal axis).

worst results in the summary–expert summaries method ($p < .05$).

Descriptive results show that reliability of the best-dimension algorithm is larger than .6 with human graders in both methods (summary–text and summary–expert summaries; see Table 1), and of the semantic common network is also larger than .6 in the summary–expert summaries method. In both cases, the semantic similarity was well in line with human assessment, and higher than the results obtained in baseline.

This table also reflects the fact that the Euclidean distance algorithm had the lowest reliabilities compared with other algorithms in the expert summary method.

DISCUSSION

We have reviewed LSA as a technique that may help us solve some positioning problems we are bound to encounter when we work with summaries. In particular, we had three main aims: to show the problem of working with very short summaries (25–50 words), to improve upon the previous study in using expository texts, and to find methods closer to human cognitive dynamic processing of texts than the standard LSA, which gives a static representation of semantics.

Concerning short summaries, it is well known that LSA has problems dealing with texts shorter than 200 words (i.e., at the sentence level, LSA results are poorer than at paragraph level; Rehder et al., 1998; Wade-Stein & Kintsch, 2004; Wiemer-Hastings et al., 1999). Our purpose in this article was to test the reliability of LSA as a computer-based procedure for assessing short summaries. In our previous study (León et al., 2006) we obtained positive results in terms of reliability between LSA and human graders, but far from those reliabilities within graders (E. Kintsch et al., 2000). This limitation is marked for expository texts (León et al., 2006). Some of LSA's limitations can be explained by the fact that LSA is not a theory of language processing. It is just a static representation of semantic memory (Jorge-Botana, Olmos, & León, in press; W. Kintsch, 2007). If we want to simulate comprehension processes with LSA, we have to create an extended semantic representation and exploit it in the way people do: for example, as people read paragraphs (Denhière et al., 2007), process entire sentences (W. Kintsch, 2007), reason (Quesada, Kintsch, & Gómez, 2005), or understand predicative metaphors (W. Kintsch, 2000; W. Kintsch & Bowles, 2002). For this reason, it is important to be very clear what LSA is and how to achieve promising results in the cognition-simulation field. The goal is to get good

Table 1
Reliability Means Between LSA and Human Graders in Each Method and Algorithm

	Algorithm				
Method	Semantic Common Network	Best Dimensions	Euclidean Distance	Baseline	Total
Summary–text method	.57	.60	.56	.53	.56
Summary–expert summaries	.61	.66	.43	.57	.57
Total	.58	.63	.49	.55	.56

results from flexible semantic representations adapted to a discourse context such as expository texts.

In order to improve LSA performance, some authors have proposed a few extensions of it. Some of these extensions have concentrated on algorithms that select dimensional information intelligently (Hu et al., 2007), or on algorithms that change the static semantic representation in a context-dependent representation (W. Kintsch, 2008). Our solution has been to create three new algorithms that incorporate (or remove) some adaptive information in the vector representation of the text. Thus, the best-dimension algorithm suppresses those dimensions that affect reliability. The common semantic network adds to the vector summary semantically related neighbors that are at the same time related to the summarized text; Euclidean distance incorporates vector length as a measure.

In overall terms, the data from our study support the reliability of LSA (using best-dimension and semantic-common network algorithms) as a tool for comparing semantic similarity with human judgment in summarizing expository text. Furthermore, LSA is able to make similar evaluations of summaries, even though we used summaries as short as 50 words in length.

Are these algorithms capable of improving assessment quality in expository text? Of the three algorithms used in this study, only the best-dimension algorithm supports this idea. In LSA, dimensions have no explicit interpretation (which is not the case for factorial analysis), but not all the semantic dimensions are task relevant. In the same way, we probably do not use all our semantic memory when we undertake a task. This algorithm seems to remove some dimensions that contain noise in assessing written material, and retain those dimensions that discriminate clearly between good and bad summaries. In the near future, the next step could be rotating the semantic space to find a new base with meaningful dimensions (Hu et al., 2007). The semantic common network algorithm also showed good tendencies and had hopeful results; it is based on the idea of simulating a number of semantic phenomena, one of which is context dependency in similarity assessment (others mentioned by W. Kintsch, 2007, are metaphor comprehension or causal inferences). This algorithm enriches vector summary with relevant terms. However, in the future it should be refined—for example, by enriching the summary only if its neighbors go over a fixed threshold. Thus, an anomalous summary might not benefit from the algorithm, and only the good (semantic) summaries would take advantage of it. Euclidean distance has not obtained enough LSA–human grader reliability in the summary–expert summaries method, but since it incorporates vector length, it would be a good measure in certain tasks (e.g., we have recently seen that the Euclidean distance algorithm can distinguish better between expert and novice answers than the cosine can). We think this method was inappropriate in the summary–expert summaries method, because the Euclidean distance is not a good enough measure, in the sense that it cannot discriminate well between good and bad summaries. In this method, LSA compares the Euclidean distance between an expert summary and a student summary, when neither can exceed the maximum of 50 words. Probably the measure of the distance between expert and student summaries is not sufficiently sensitive and cannot provide good ratings, since both are forced to not exceed a maximum length; we think this is the most plausible explanation for the variations of behavior (interaction) we have found with this algorithm. This method works well when at least some of the texts compared are not limited in length, as we can see in the summary–text method.

Future research will have to incorporate new ideas to confront the new task demands—ideas that link LSA with psychological models like W. Kintsch's (2001), or ideas that improve its mathematical possibilities, as have been proposed by Hu et al. (2007). For instance, W. Kintsch (2001) has shown that algorithms beyond those that compute the overall similarity of one sentence to another are needed for LSA to account for language use in which the interpretation of arguments is dictated by the context, as in metaphor comprehension, causal reasoning, and some similarity judgments. One of the main aims of applying LSA successfully is to understand the psychological processes of interest and how semantic and dynamic relatedness might impact those processes but not always come up with convincing results. For example, if students are instructed to read and summarize a text, we would expect a high degree of similarity between the meaning of the text and the students' summaries. But if students were asked to write essays, instead of summaries, applying what they had read to a new situation, the resulting essays might be expected to be less similar than the summaries to the original text. In addition, we would expect less similarity across students, and we might not expect the level of similarity to predict essay quality. In other words, LSA may not be appropriate for analyses of reasoning phenomena; or, at least, simple similarity indices may not be. The ability of LSA to match human semantic-relatedness judgments is dependent on LSA's having exposure to texts comparable to the ones human judgment makers would have had exposure to. Further research should use these or similar algorithms in other types of text and other types of task as a way to validate and generalize previous research.

AUTHOR NOTE

This work was supported by Grant SEJ2006-09916 from the Spanish Ministry of Science and Technology. We thank anonymous reviewers for their helpful comments. Correspondence concerning this article should be addressed to J. A. León, Facultad de Psicología, Universidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain (e-mail: joseantonio.leon@uam.es).

REFERENCES

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, 391-407.

Denhière, G., Lemaire, B., Bellissens, C., & Jhean-Larose, S. (2007). A semantic space modeling children's semantic memory. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 143-167). Mahwah, NJ: Erlbaum.

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

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A. C., & McNamara, D. S. (2007). Strengths, limitations, and extensions of LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 401-426). Mahwah, NJ: Erlbaum.

Jorge-Botana, G., Olmos, R., & León, J. A. (in press). Using LSA and predication algorithm to improve sense extraction in a diagnosis Spanish corpus. *Spanish Journal of Psychology*.

Kintsch, E., Steinhart, D., Stahl, G., & LSA Research Group (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, **8**, 87-109.

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

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

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157-170). Amsterdam: Benjamins.

Kintsch, W. (2007). Meaning in context. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.) *The handbook of latent semantic analysis* (pp. 89-105). Mahwah, NJ: Erlbaum.

Kintsch, W. (2008). Symbol systems and perceptual representations. In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 145-164). Oxford: Oxford University Press.

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor & Symbol*, **17**, 249-262.

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 the 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.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, **38**, 616-627.

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.

McNamara, D., Boonthum, C., Levinstein, I., & Millis, K. K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

Quesada, J. F., Kintsch, W., & Gómez, E. (2005). *Latent problem solving analysis (LPSA): A theory of representation in complex problem solving.* Manuscript submitted for publication.

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

Wade-Stein, D., & Kintsch, E. (2004). Summary street: Interactive computer support for writing. *Cognition & Instruction*, **22**, 333-362.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). How latent is latent semantic analysis? In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence* (pp. 932-937). San Francisco: Morgan Kaufmann.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In M. Danson (Ed.), *Proceedings of the 9th International Computer Assisted Assessment Conference* (pp. 485-494), Loughborough, U.K.: University of Loughborough.

(Manuscript received August 29, 2008;
revision accepted for publication March 20, 2009.)