Using Latent Semantic Analysis to grade brief summaries: A study exploring texts at different academic levels.

Ricardo Olmos & José A. León

Universidad Autónoma de Madrid, Spain

&

Guillermo Jorge-Botana & Inmaculada Escudero

Universidad Nacional de Educación a Distancia

Running Head: LSA to grade summaries

Please send correspondence to:

Ricardo Olmos Albacete

Departamento de psicología social y metodología

Facultad de Psicología

Universidad Autónoma de Madrid

Campus de Cantoblanco

28049-Madrid, Spain

e-mail: ricardo.olmos@uam.es

Telephone: +34 91 497 85 86

Abstract

In this study we propose an integrated method to automatically evaluate very brief summaries (around 50 words) using the computational tool Latent Semantic Analysis (LSA). The method proposed is based on a regression equation calculated with a corpus of a hundred summaries (the training sample), and is validated on a different sample of summaries (validation sample). The equation incorporates two parameters extracted from LSA: (1) the semantic similarity of the summary, measured using the Summary–expert summaries method (Landauer et al., 1998; León et al., 2006; Olmos et al., 2009) and (2) the vector length (Redher et al., 1998). The study is based on a sample of 786 summaries by students at four academic levels. All of these students summarized either an expository or a narrative text; their summaries were then evaluated by four graders on a scale of 0-10. The results support three ideas. First, that incorporating both parameters into the method is more successful than the traditional cosine measure. The reliability of LSA for evaluating summaries rises above the 0.80 level for the expository text. Second, that LSA shows practically the same level of sensitivity as the human graders to the quality of the summaries at different academic levels. Third, that the method overcomes a serious limitation of LSA: its difficulties evaluating very brief texts (Redher et al., 1998; Wiemer-Hastings et al., 1999).

Keywords: Latent Semantic Analysis, assessment summaries, academic levels, university students, Secondary students, Primary students, vector length, expository text, narrative text.

Introduction

In recent decades educators have often used multiple choice tests to evaluate comprehension of material. There are undoubtedly advantages to multiple choice testing, including speed of assessment and the possibility of evaluating many different aspects in a short time-frame, low cost, objective reliability measures (test-retest, Cronbach's alpha, etc.) and relatively simple analysis of the psychometric properties of items. This form of assessment has its limitations, however: comprehension would be more superficial than for a student challenged with an open-ended question (Millis, Magliano, Wiemer-Hastings, Todaro & McNamara, 2007; Shapiro & McNamara, 2000). The cognitive demands of a multiple choice test are those of recognition rather than recall, so the student's learning strategy does not demand deep understanding of the text (Far, Pritchard & Smitten, 1990). Causal relationships, drawing of inferences and the form of expressing ideas would not be assessed by this type of tasks.

Exposing students to open-ended tasks, on the other hand, offers a means of evaluating deeper understanding of the material. From the constructivist perspective, building explanations or producing written material such as summaries gives rise to and improves comprehension of texts more than multiple choice testing (Graesser, Lu, Jackson, Mitchell, Ventura, Olney & Lowerse, 2004). Several studies have demonstrated the importance of knowing how to summarize succinctly in understanding and learning, and have shown it is a good measure of comprehension processes (Brown, Bransford, Ferrara & Campione, 1983; Wade-Stein & E. Kintsch, 2004).

Summaries play a major role in research on comprehension of texts and its assessment. For some authors a text has not been understood if the reader cannot summarize it (Palinscar & Brown, 1984). However, it is often taken for granted that students learn to summarize as they move to higher-level studies. In countries like Spain, however, courses teaching summarizing skills are uncommon. Whilst it is true that the actual concept of a summary is somewhat imprecise, there might be general agreement that the process of producing a good summary implies understanding a text, identifying the main ideas and transmitting them succinctly (E. Kintsch, Caccamise, Franzke, Johnson & Dooley, 2007; León et al., 2006). For authors such as van Dijk and Kintsch (1983) summarizing involves the capacity to generalize, synthesize and write coherently. It thus goes far beyond reading, since it implies profound comprehension of what is read. In their model of comprehension, summarizing is essential to understanding, since it involves extracting the main content of what is read, and at the same time eliminating superficial details. Kintsch himself (2002) used the LSA model in an attempt to find the phrases that best summarize the content of a text. To achieve this aim he sought structures (titles, subheadings or paragraphs) that best represent the information contained in the text (the abstract information from the macrostructure).

Summarizing, then, involves establishing relationships between important concepts, and presenting them in a coherent, organized manner. The information must be restructured, further abstracting it from the content of the text. The summary allows us easier access to factual and conceptual knowledge in memory. Summarized texts allow us to build on information in the classroom much more and better than by simply rereading a text. It allows students to formulate more pertinent questions, and the teacher to evaluate the extent to which the material was understood. In this line we subscribe to a popular eighties school of thought (see review by Bransford, Brown & Cocking, 2000) that learning to summarize is a central aspect of the comprehension process, so that reliably evaluating a summary is key to knowing whether a student has a deep understanding of a text.

The demands placed on teachers make it very difficult to find time to evaluate essays and summaries and give feedback to each of the students individually; so many educators favor multiple choice tests (Wade-Stein & E. Kintsch, 2004). There are tools currently available which can evaluate texts reliably (see review by Dikli, 2006). These tools, although they are not apt for awarding a final grade, might be valuable for monitoring a student's progress, or to provide the student with longitudinal information regarding their level of ability. The task of introducing these tools in the classroom is extremely complex, and depends on many factors (e.g. schools infrastructure, budget, subject matter and access to technology) and also on cooperation between many individuals (e.g. educational psychologists and teachers). However, one of these contributions - Automated Essay Scoring - is already far advanced enough to take a step forward.

This paper was driven by two factors: on the one hand the need to introduce summaries into the classroom to synthesize the key information from syllabus units covered, and on the other the possibility of extracting reliable automated evaluations of these summaries using a computational tool known as *Latent Semantic Analysis* (LSA). These two aspects have normally been impossible to reconcile, since LSA only provides reliable evaluations using more extensive texts - mostly over 250 words (Rehder, Schreiner, Wolfe, Laham, Landauer & Kintsch, 1998). In fact, some authors indicate that LSA has special difficulty in analyzing texts between two and sixty words (Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999) - the range of the summaries used in our research. In this study, then, we propose an LSA-based method that provides reliable evaluations of very brief summaries. The method we propose combines information on semantic similarity commonly provided by LSA, together with information on the extent of knowledge LSA has of the terms represented in a semantic space. This means that the method combines the cosine measure with vector length information. We will begin by looking at the importance of summaries in the classroom together with the need for evaluation of open-ended responses as a complement to multiple choice exams. In fact, this paper attempts to analyze a combination of ratings from LSA cosine and vector length measures on the one hand, and human graders' evaluations of content and coherence on the other. These types of measures will be applied in the assessment of brief summaries by students at different academic levels and using two types of texts (a narrative text and an expository text). In terms of academic levels, we took 238 students from 6^{th} grade, 192 students from 8^{th} grade, 198 students from 10^{th} grade, and lastly 158 university students. In addition, these open-ended responses will be compared with standard multiple choice scores obtained by the same students on the same texts.

LSA in educational tasks

Latent Semantic Analysis (LSA) is a computational technique that contains a mathematical representation of language. During the last twenty years its capacity to simulate aspects of human semantics has been widely demonstrated (Landauer & Dumais, 1997). LSA is based on three fundamental ideas (Steyvers & Griffiths, 2007):

(1) To begin to simulate human semantics of language we first obtain an occurrence matrix of terms by document, (2) the dimensionality of this matrix is reduced using singular value decomposition (SVD), a mathematical technique that effectively makes the tool a latent semantic space, and (3) any word or text is represented by a vector in this new latent semantic space. Since each word is a vector, a text is the sum of the words that comprise it, i.e. another vector.

To evaluate the semantic similarity between two texts, we need only extract the cosine formed between the two vectors. When the cosine is close to zero the semantic similarity is null. When the cosine is close to one the semantic similarity is very high. Formally, the cosine is defined as:

$$Cos = \frac{\sqrt{\sum_{i=1}^{k} x_i y_i}}{\sqrt{\sum_{i=1}^{k} x_i^2}\sqrt{\sum_{i=1}^{k} y_i^2}}$$

where the numerator contains the scalar product between the *k* coordinates of the first and second word (*x* and *y* respectively), and the denominator contains the product of the word *x* and word *y* vector lengths.

There are other possible LSA-derived measures that could be used, such as the dot product or Euclidean distance between the two vectors, or the length of an individual vector. A vector can be thought of as a position within an n-dimensional space. The value of a vector is represented as a series of coefficients, each coefficient representing a value (or distance) along a particular dimension in the n-dimensional space (Reder, Schreiner, Wolfe, Laham, Landauer & Kintsch, 1998). The vector length formula can be defined as:

$$VectorLength = \sqrt{\sum_{i=1}^{k} x_i^2}$$

that is, the square root of the sum of squares of the *k* coordinates that represent the word *x*.

The vector length also informs us of the knowledge LSA has of this text. A graphical representation of how LSA works can be seen in Figure 1. There are several points to note regarding the diagram. First, that if we semantically compare the text “lush forest” with “tropical jungle” the cosine is close to one, since both vectors have very similar directions due to their similar meanings. However, these two vectors have a cosine close to zero with the text “modern building”, since the vectors produce a practically orthogonal angle. Second, since the vector length for “modern building” is greater than that of the other two vectors, we can assume that LSA has more knowledge about this text than about the other two. In short, after LSA creates a semantic space in which it can represent the texts vectorially, we can calculate the semantic relationship between two texts using the cosine formed by the two vectors, and measure the knowledge that LSA has of a text by the length of the vector that represents it. Vector length contains the quantity of summary elaboration, and the cosine contains the quantity of semantic similarity.

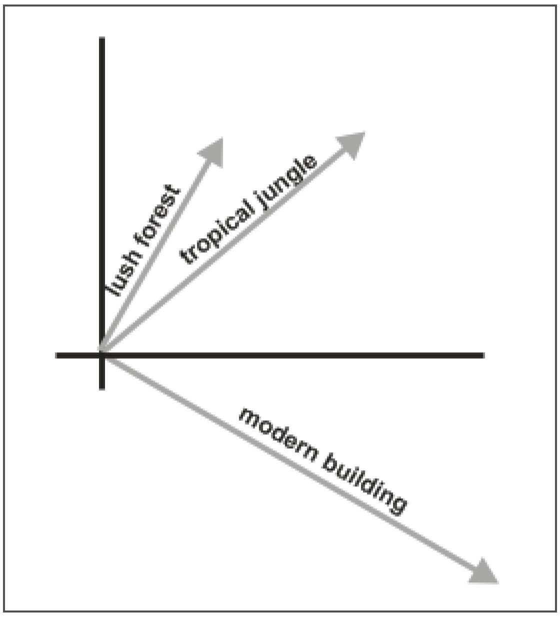

Figure 1. Graphical example of LSA representing three texts (vectors)

LSA assessment methods normally only use the former to attempt evaluations, ignoring vector length (Foltz et al., 1999; Landauer & Dumais, 1997; Landauer et al., 1998; Wade-Stein & Kintsch, 2004). Nonetheless, some studies have analyzed vector length in the assessment. Redher et al. (1998) used a multiple regression to demonstrate that the combination of the semantic similarity (cosine) and vector length explain the correlation of LSA with human criteria, and that in fact vector length is the factor responsible for the greater part of the variance. In this study the authors examine these and other measures derived from the cosine or the vector length, such as the Euclidean distance or the scalar product, to test their efficiency in automatic evaluation of essays. They found that in the regression the strongest predictors were vector length and cosine, although alone (i.e. not in combination with the others) the most powerful measure was the scalar product. Other research has also covered alternative measures such as Euclidean distance in the automatic grading of summaries (Jorge-Botana, León, Olmos & Escudero, 2010; Olmos, León, Jorge-Botana & Escudero, 2009).

LSA has been used abundantly in the field of education. For example, it has been used to evaluate online comprehension using verbal protocols while asking students to read (usually self-explanations) (McNamara, Levinstein & Boonthum, 2004; Millis et al., 2007). LSA is a satisfactory tool for capturing the meaning of these verbal protocols. In these studies LSA reveals different kinds of learning strategies when the students read, or rather when they discuss what they are reading. It detects, for example, that some tend to paraphrase what they just read, whilst others usually relate it to other phrases read previously, or to their prior knowledge. Since these different strategies generally imply different levels of comprehension, LSA can be used quite successfully either to predict the level of comprehension, or to evaluate the predominance of reading strategies or give appropriate feedback to students, coaching them towards better use of strategies (McNamara, Boonthum, Levinstein & Millis, 2007; Millis et al., 2007).

Another educational application of LSA is with computer tutors (Graesser, Chipman, Haynes & Olney, 2005; Wade-Stein, E. Kintsch, 2004). Graesser et al. (2005) created a tool called AutoTutor, using a computer to hold conversations with students in natural language. Students are presented with common problems from the curriculum script, and using an animated agent AutoTutor gives them feedback until they manage to give a satisfactory response to each problem. Another computer tutor is Summary Street by Wade-Stein and E. Kintsch (2004). This tool coaches during the process of writing a summary. The basic idea underlying this tool is how to teach students to summarize. These computer tutors teach students to resolve problems (e.g. summarizing) without individual attention from a teacher - something totally impossible if we consider the lack of educational resources available today. In simple terms, what

LSA does is compare what students write with texts incorporated into the tools (ideal summaries, main topics, key words, etc.) using the cosine measure. Thresholds are set such that if the cosines rise above them we assume the student response covers the pertinent aspects. If the cosine does not reach the threshold, the computer tutor gives clues to help the student incorporate the missing information. The central feature is the dynamic interaction between student and machine. If a stimulating environment is combined with good task design, the results show a notable improvement in student responses (Graesser, Penumatsa, Ventura, Cai & Hu, 2007; E. Kintsch et al., 2007).

A third example of an LSA application in the field of education is based on automatic assessors (Foltz, Laham & Landauer, 1999; Landauer, Foltz & Laham, 1998). For example, Foltz et al. (1999) created the Intelligent Essay Assessor (IEA). These methods are based on using the cosine to compare student essays with a source text. One very common method is when the source text consists of an expert summary (normally by a grader or teacher), thus creating what they call a "golden summary" (Landauer et al., 1998; León et al., 2006). These tools automatically provide an essay score, sometimes offering impressive results (e.g. above .80), proving as reliable as the expert judges (trained graders or teachers) themselves. Another similar application by the French authors Dessus & Lemaire is the APEX system (2002). This system, based on LSA, provides texts for students to summarize, and then evaluates the summary. Other more innovative procedures include the so-called EssayAid (Kakkonen & Sutinen, 2011) for semi-automatic essay evaluation, or evaluation procedures that combine LSA with n-grams (Monjurul Islam & Latiful Hoque, 2012).

More recent studies have introduced the possibility of evaluating bridging inferences based on text cohesion, measuring this cohesion as the similarity between adjacent phrases using the cosine with LSA. The tool is known as Interactive Strategy Trainer for Active Reading and Thinking (iSTART) created by Bellissens, Jeuniaux, Duran McNamara (2010). This tool was conceived so that with the help of LSA students can improve their reading strategies. In fact they promote what they call Self-Explanation Reading Training (SERT).

Lastly, we do not want to give the impression that LSA is only applied to the evaluation of essays. Its versatility makes it a fertile resource in more complex tasks - for example LSA can be used to adapt texts to the participant's level of ability, so that they make the best use of the texts they learn with (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch & Landauer, 1998). LSA is not only applicable to the field of written tasks, as we see for example in its successful application to reasoning task analysis for complex problems (Quesada, Kintsch & Gomez, 2001).

This study focuses on the latter type of applications, more specifically on presenting a method that allows us to evaluate summaries reliably. LSA-based evaluations have normally been applied to relatively long essays (over 200 words), but few have tackled the use of LSA to evaluate brief summaries of only fifty words. When texts contain fewer than 1000 words, it is only natural that the summaries are shorter than those used in most LSA applications (Wade-Stein & Kintsch, 2004; Landauer et al., 1998, Rehder et al., 1998). For a detailed up-to-date review of research on automatic essay evaluation with LSA see Haley (2009). In our case, the mean number of words across the 396 summaries of a narrative text (which contained 402 words) was 39.15 – a ratio of 1 word in the summary for every 10 in the text. The mean number of words in the 388 summaries of an expositive text (500 words) which was also applied was 28.22, giving a ratio of approximately 1 word in the summary for every 18 in the text.

Objectives

The aim of this study was to use a LSA-based computational method to reliably evaluate especially brief summaries (maximum fifty words). The method incorporates the essential information from the latent semantic space: (1) A measure of semantic similarity using the cosine and (2) a measure of the vector length or extent of knowledge about the text. With this general aim we sought four goals. First, to try to obtain reliable evaluations (at least > 0.70) combining both kinds of essential information derived from the latent semantic space (cosine and vector length, analyzing their effects separately as well as in combination) compared to human graders and for each type of text. Second, to take Pearson correlations between LSA measures (cosine and vector length) vs. human graders' measures (content and coherence), related to summary length (number of words). Third, to analyze whether LSA is sensitive to the quality of brief summaries by students at different academic levels, as trained human graders are. And fourth, to analyze whether positive correlations exist between summaries and multiple choice scores.

Method

Participants. 786 students from four grade levels took part in this study. Student ages ranged from 10 to 23 years. The youngest students were from 6th grade (a total of 238 students), followed by 192 students from 8th grade, 198 students from 10th grade, and lastly 158 second year psychology undergraduates.

Material. Two texts were applied in this study. A narrative text (*The Carob Tree Legend*) was 402 words long and its comprehension requires only general knowledge. The expository text (*The Strangler Trees*) was 500 words long and understanding it also requires general knowledge.

The Spanish LSA corpus. The generalist corpus used in this study belongs to the University of Colorado. The corpus contains material from on-line encyclopedias, newspapers, textbooks and several Internet sources. In total the corpus has 2,059,234 documents (i.e. paragraphs) and 1,661,954 different terms. A semantic space with 337 dimensions was used. The method used was document to document.

Procedure. Each student read the text at their own pace in a classroom. Before reading the text they were told that it was important to understand the text in order to answer a set of questions. After reading, students were allowed 15 minutes to write a summary (maximum fifty words). Finally, they were asked to answer a set of multiple choice comprehension questions. The 786 summaries used in this research, 396 of them were summaries of a narrative text, and the remaining 390 were summaries of an expository text.

The four judges' evaluations. To obtain the expert judges' evaluations with which to later compare those of LSA, four doctorate students received four sessions of instruction in evaluating summaries on a scale from 0 to 10. The evaluation criteria followed were taken from research by León & the Reading Literacy Research Group. Whilst grading the summaries, the judges took two aspects into consideration. First the content of the summaries was evaluated on a scale of 0 (no content) to 4 (all key content). Both the narrative and the expository texts contained four main ideas that had to be considered in the summaries (see León et al., 2006). Each main idea counted as one point. Secondly coherence was evaluated on a scale of 0 (incoherent) to 6 (highly coherent). To assess coherence the organization, causal relationships, use of connectives, extent of conceptualization of the summary and lack of redundancy were analyzed. The judges carried out the evaluations independently and without knowing the student's academic level.

The proposed method

To implement our method we used a database comprising 107 summaries of narrative text and 93 expositive summaries, distributed across four grade levels. The sample used to adjust the method is called the training sample, and allows us to calculate the way we obtain the scores with LSA, although it is not used to evaluate the reliability of the method. Table 1 shows scores from the 200 summaries in the training sample used to implement the method.

		Educational level				
		6th grade	8th grade	10th grade	University	Total
Type of text	Narrative	27	29	32	19	107
	Expository	29	26	24	14	93
	Total	56	55	56	33	200

Table 1. Training sample of summaries in the narrative and expository text

Each of these summaries was graded independently by each of the four judges on a scale of 0 to 10, awarding up to four points for content and six points for coherence of the summary. Blind scoring was used, in other words the graders were unaware of the student's academic level. An average score was obtained using the four graders' scores. After this a regression line was calculated, where the dependent variable was the graders' average score and the independent variables were vector length and semantic similarity.

The regression equation for the narrative text was:

$$NarrativeScore = \beta_0 + \beta_1 * Vectorlength + \beta_2 * Similarity$$

And the regression equation for the expository text was:

$$ExpositoryScore = \beta_0 + \beta_1 * Vectorlength + \beta_2 * Similarity$$

where β_0 is the constant, β_1 is the coefficient for vector length and β_2 is the coefficient for semantic similarity.

Once the regression calculations are done, we took a new sample of summaries (the validation sample). This time there were 289 summaries of the narrative text and 297 summaries of the expositive text. Table 2 shows the distribution of summaries for the validation sample used to check the method.

		Educational level				
		6th grade	8th grade	10th grade	University	Total
Type of text	Narrative	92	69	68	60	289
	Expository	90	68	74	65	297
	Total	182	137	142	125	586

Table 2. Validation sample of summaries of the narrative and expository text

Summaries were again graded independently by the four graders, using the same scale from 0 to 10. An average was taken of the four judges' scores to obtain a single grade. These grades were used to assess the reliability of the scores awarded by LSA using the regression equations. The independent validation sample was used to avoid overfitting, to make it easier to generalize to new summaries.

The two LSA measures, vector length and semantic similarity (cosine), were obtained as follows. Given that in LSA each document is represented by a vector, the vector length is simply calculated as the length of each summary vector. In the equation the vector length component represents how detailed the summary is - the greater the vector length the more detail, and the more familiar or relevant words appear in the semantic space. The measure of similarity is somewhat more difficult to obtain. For this reason we use a well-known method, habitually used in Automated Essay Scoring with LSA: the Summary–expert summaries method (Dikli, 2006; Foltz et al., 1999; Kintsch et al., 2007; León et al., 2006; Landauer & Dumais, 1997; Landauer et al., 1998; Olmos, et al., 2009). The Summary–expert summaries method consists 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 semantically similar 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 scores a student summary as follows: LSA computes cosines between the student summary and each of the six expert summaries. The final score for the student summary is the average of these six cosines.

To obtain a measure of similarity we had to compare the summary with a source text. This method uses ideal summaries, in other words summaries written by experts containing the essential information from the text (all relevant details and very strong coherence). To this end, six teachers with expertise in comprehension were asked to write a summary of no more than fifty words, both of the expository text and of the narrative text. To obtain a measure of semantic similarity the cosine between the student summary and each expert summary was calculated. Since there are six cosines, one for each expert summary, the average cosine is taken as the final measure of semantic similarity. Both vector length and semantic similarity values were obtained automatically for all 786 summaries, as described. What we refer to in the regression equation as *similarity* will be called the *Expert Method* henceforth, since the similarity is calculated using the expert’s method.

Data analysis. The data was analyzed in four stages. First the regression equation with the ordinary least square estimation method that best fits the judges' scores was obtained using the training sample, as described in the previous section. Secondly, the reliability (Pearson correlations) between LSA and judges was calculated using the validation sample. Third, an ANOVA was used to study the sensitivity of judges and LSA to differences in the quality of summaries from different grade levels. Lastly, the reliability (Pearson correlations) between LSA measures (cosine and vector length) vs.

human graders' measures (content and coherence), related to summary length (number of words), and the relationship between summary and multiple choice scores.

Results

Regression equations

To predict the average judges' scores a training sample was used, comprising 107 summaries of the narrative text and 93 summaries of the expository text chosen completely at random. These samples were used to find two regression equations, later applied to predict the judges' grades for the validation sample. The regression line to predict the grades for the summaries of the narrative text was:

$$NarrativeScore = -1.62 + 5.76 * Vectorlength + 11.26 * ExpertMethod$$

Where the coefficient of Vector Length was statistically significant ($t(1)=4.98$, $p < .05$), as too was the Expert Method coefficient of semantic similarity ($t(1)=6.34$, $p < .05$). The tolerance between the two variables was .80.

The regression line obtained for the expository text was:

$$ExpositoryScore = -4.19 + 10.18 * Vectorlength + 15.61 * ExpertMethod$$

Once again, the coefficient of Vector Length was statistically significant ($t(1)=8.22$, $p < .05$) as was the coefficient associated with Expert Method ($t(1)=8.63$, $p < .05$). The tolerance between the two variables was .92.

To obtain a grade for a new summary (narrative or expository), then, we had only to calculate the vector length of the summary and a measure of semantic similarity using the expert method.

We then substitute into the corresponding equation to obtain the grade for the summary. Both equations obtained positive coefficients, informing us that the greater the summary detail (as measured by vector length) the greater the summary grade, and the greater the semantic similarity between the student summary and the experts' summaries the greater the summary grade.

Lastly, we examined the impact of vector length on reliability, compared to the traditional method using only the measure of similarity (Expert method) (Foltz et al., 1999; Landauer et al., 1998; Wade-Stein & E. Kintsch, 2004). It was found that on including the vector length information, the change in the proportion of variance explained was statistically significant in the case of the narrative summaries ($F(1,104) = 24.81$, $p < .05$) as well as in the case of the expository summaries ($F(1,90) = 67.51$, $p < .05$). In other words, with these brief summaries there was a substantial drop in reliability if we do not incorporate vector length (see following section).

A) Reliability between the four human graders themselves

Before analyzing the reliability of LSA, we evaluated reliability between the four judges themselves. This reliability varied between .78 and .86 for the narrative text, and between .83 and .88 for the expository text - all highly reliable and statistically significant.

B) **Reliability between LSA and human graders' scores.**

The LSA-grader reliabilities were calculated using the validation sample. This sample was set aside to avoid overfitting of reliabilities, and thus allow generalization of results to other summaries. The validation sample consisted of 289 summaries of the narrative text and 297 of the expository text. The LSA grades for this new sample of summaries were obtained by simply using the equations presented in the previous section.

Table 3 shows the reliability of LSA with individual judges and with the average judges' scores in both texts (calculated with Pearson's correlation). For the narrative text the reliability of LSA ranged from .60 to .67 for the individual judges, and reached .68 with the average judges' scores. As for the expository text, the LSA-grader reliability ranged from .76 to .78, reaching .82 with the average judges' scores. For the narrative text the reliability scores were fairly high; in the expository text they were high.

	Human Grader				
Text	Grader 1	Grader 2	Grader 3	Grader 4	Average grader
Narrative text	$.61^{**}$	$.67^{**}$	$.60^{**}$	$.63^{**}$	$.68^{**}$
Expository text	$.76^{**}$	$.78^{**}$	$.77^{**}$	$.78^{**}$	$.82^{**}$

** $p<.01$.

Table 3. LSA-grader reliability for each text and human grader

As for the LSA-grader reliability comparison we found significant differences between LSA-human graders' reliability for the expository text ($r = .82$, $p<.01$) and LSA-human graders' reliability for the narrative text ($r = .68$, $p<.01$). Reliability for the expository text was thus significantly higher ($p<.01$).

Figures 2 and 3 show *scatter plots*, as a graphical representation of LSA evaluations compared to the average judges' grades. Each point represents a pair of grades: the judges' average and the LSA grade.

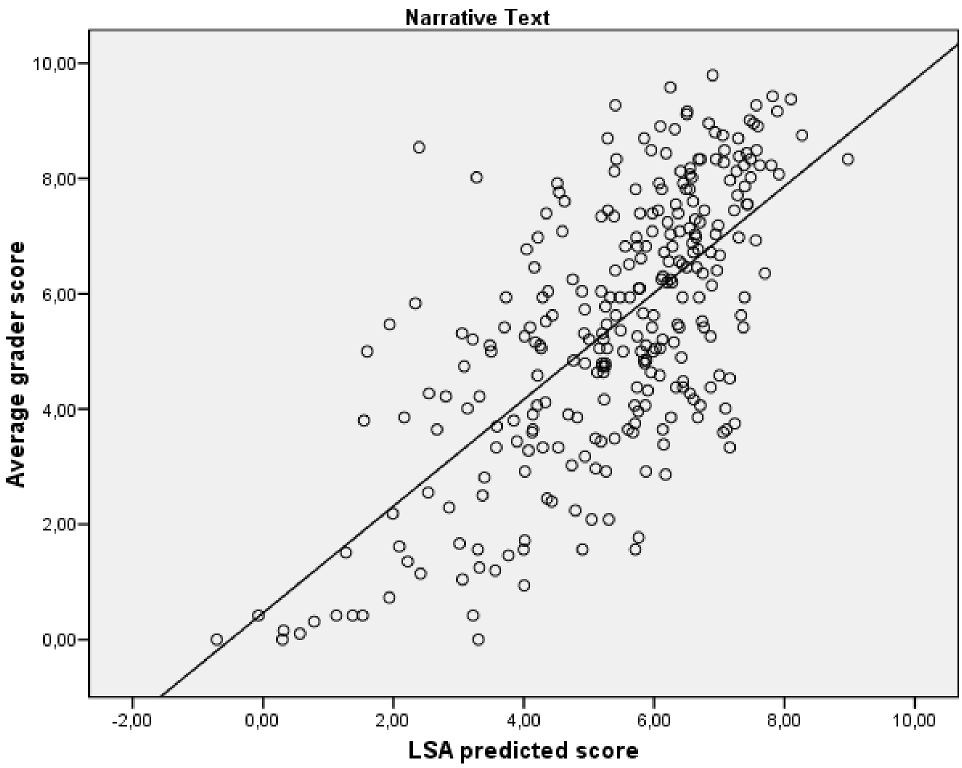

Figure 2. Average grader scores plotted against LSA-predicted narrative scores.

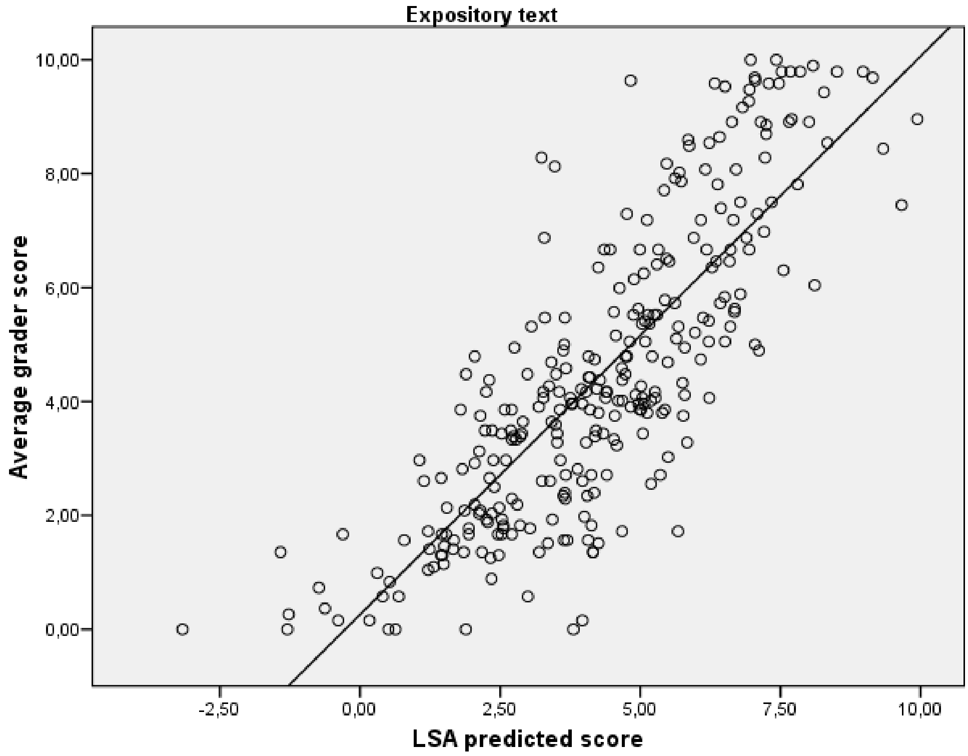

Figure 3. Average grader scores plotted against LSA-predicted expository scores.

The graphs show the average judges' scores against the scores predicted by LSA using the regression equation above. As we saw from the table of reliability scores, the cluster of points fits the line better for summaries of the expository text than for those of the narrative text. It is clear that when the judges award the summary a low score LSA does the same. LSA also follows suit when the judges give a high grade. Both graphs show linear relationships; there were no summaries whose LSA grade differed markedly from the average judges' grade.

C) **Pearson correlations between LSA measures (cosine and vector length) vs. human graders measures (content and coherence)**

In Table 4 we can see the data analyzed by type of measure, in order to evaluate the results of the previous section in more detail.

		Judge's assessments		
	LSA	Content	Coherence	Global (content + coherence)
Narrative summaries	Cosine (expert method)	.61**	.51**	60**
	Vector length	.56**	.46**	.55**
	Cosine + Vector length	.68**	.56**	.68**
Expository summaries	Cosine (expert method)	.68**	.63**	.67**
	Vector length	.64**	.65**	.66**
	Cosine + Vector length	.82**	.81**	.82**

**p<.01.

Table 4. Table Correlations between LSA and Judge’s assessments for each text and type of measure (cosine, vector length, content, coherence.

Whilst the results obtained are all significant, they highlight differences that should be borne in mind when considering the type of text. For the expository text there is a high degree of homogeneity in the correlations obtained between LSA and experts, which range from .63 to .68. There is a positive effect with the combination of cosine and vector length, offering a very high correlation of .82. In contrast, the correlations obtained for the narrative text show greater divergence between LSA cosines and human evaluations of content and coherence. Whilst the cosine correlates highly with content (.61), the score for coherence is lower (.51). Vector length correlates less with both coherence (.46) and content (.56), and the combined effect of cosine and vector length does not strengthen the result of correlations with human graders. It is clear, then, that expository text summaries are better evaluated using LSA in all cases.

There were significant differences in the reliability (Pearson) between human graders (content and coherence) and LSA, but only for the narrative summaries. The combined cosine and vector length LSA method offers significantly higher correlation with human assessment of content ($r=.680$, $p<.01$) than of coherence ($r=.563$, $p<.01$). However, in the expository summaries there are no significant differences in the correlations between LSA and grading of content ($r = .823$, $p<.01$) and LSA and grading of coherence ($r = .806$, $p<.01$).

D) **Correlations between LSA measures (cosine and vector length) vs Human grader measures (Content and coherence) related to the length of summary**.

In this section we have included a new analysis of the measures made by human graders (content, coherence and the two combined) and LSA (cosine and vector length, and the two combined) regarding the length of the summary made by each student, to assess whether scores from LSA and Human graders correlated with the length the summary.

The results show a significant correlation in every case, which may indicate that summary length affects the score awarded by both expert and LSA. This result is not unexpected as both content and coherence of a summary are favored by longer summaries with more information to evaluate. The same occurs here with LSA, although the correlations are even higher when analyzed using vector length, which is more sensitive to summary length as it measures length as well as familiarity (vector length has been interpreted as the familiarity LSA has with technical words, Rehder et al., 1998).

Type of text		Number of words in the summary
	Cosine	.52**
	Vector length	.88**
	Cosine + vector length	.89**
Narrative	Content	.64**
	Coherence	.65**
	Content + coherence	.69**
	Cosine	.49**
	Vector length	.83**
Expository	Cosine + vector length	.87**
	Content	.76**
	Coherence	.78**
	Content + coherence	.79**

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

Table 5. LSA-Human grader reliability for each text and type of measure (cosine, vector length, content, coherence) related to the length of summary

In Table 5 we can see that summary length (number of words) affects the assessment of human graders whether evaluating content or coherence. The mean number of words in the narrative text summary was 39.15 (S.D.=16.28 (range=88). The mean number of words in the expositive text summary was 22.22 (S.D=14.66, range=89). Thus we could say that the longer the summary the higher the grade awarded for both content and coherence. This phenomenon is more pronounced for the expository than the narrative text. It can also be seen in LSA assessment, but only when applying vector length, which is more sensitive to summary length. This is not an unexpected result given the nature of this measure, but is not the case when cosine is used. The fact that the cosine is not susceptible to this effect may indicate that the two measures are complementary, and that together they can allow us to better evaluate a summary.

E) Sensitivity to differences between different academic levels

Lastly, an ANOVA was carried out on each text to study the capacity of both judges and LSA to detect differences in the quality of summaries from different grade levels. In this case the analysis was carried out using only the validation sample. The results obtained showed that in the narrative text judges detected more differences in the quality of summaries than LSA. For the expository text, however, the results for LSA and human graders were similar.

Figure 4 shows the average LSA and judges' grades for summaries of the narrative text at each of the four grade levels. Both LSA and the judges detect differences in the quality of summaries ($F(3,285) = 10.96$, $p < .05$ and $F(3,285) = 36.60$, $p < .05$, for LSA and average judges' scores respectively). However, a *post hoc* test revealed three groups of averages for the judges and only two groups for LSA. The judges detected that the group that summarized with the highest quality was the undergraduate group, followed by 10th grade and then 6th grade and 8th grade students (no difference was found between these last two). LSA detected two groups of averages: the group that summarized with the highest quality was the university students and the other grade levels formed another group of lower-quality summaries (no statistically significant differences were detected between them).

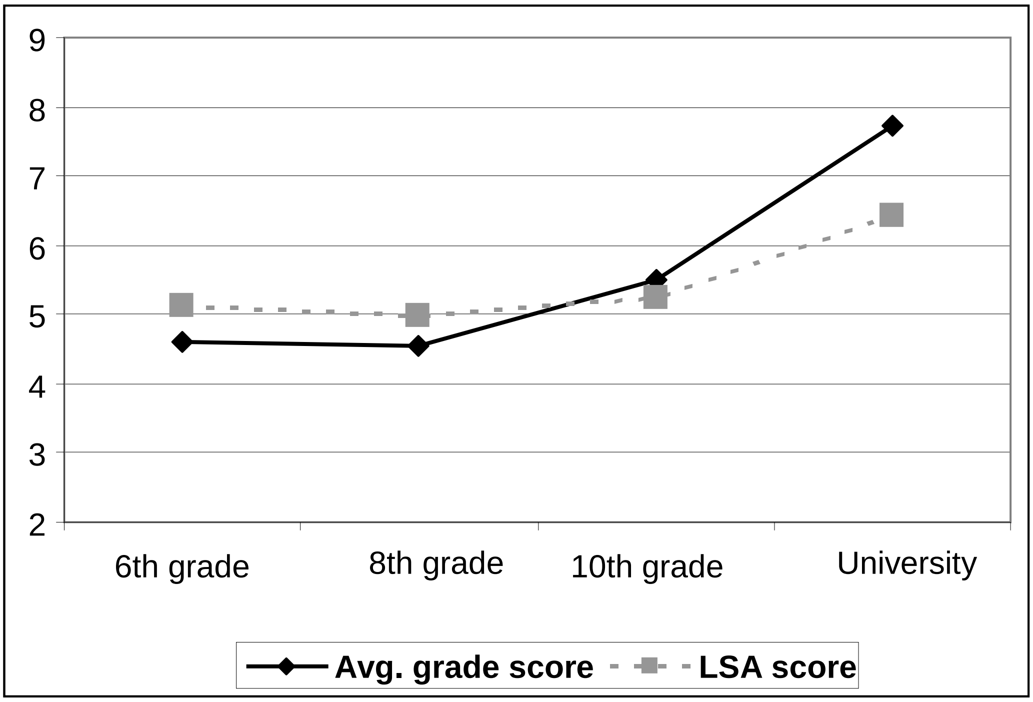

Figure 4. LSA and graders' mean score at each grade level for the narrative text

At the same time, Figure 5 shows the average LSA and judges' grades for summaries of the expository text at each of the four grade levels. As for the narrative text, both LSA and the judges detected differences in the quality of summaries ($F(3,293) = 47.88$, $p < .05$ and $F(3,293) = 89.97$, $p < .05$, for LSA and average judges' scores respectively). This time the *post hoc* test showed that both LSA and the judges detected two groups of averages: the undergraduate group summarized with higher quality than the other grade levels (no statistically significant difference was found between the averages). In Figure 5 we see the pattern of averages for LSA and judges' average scores are practically identical, corroborating the high reliability shown.

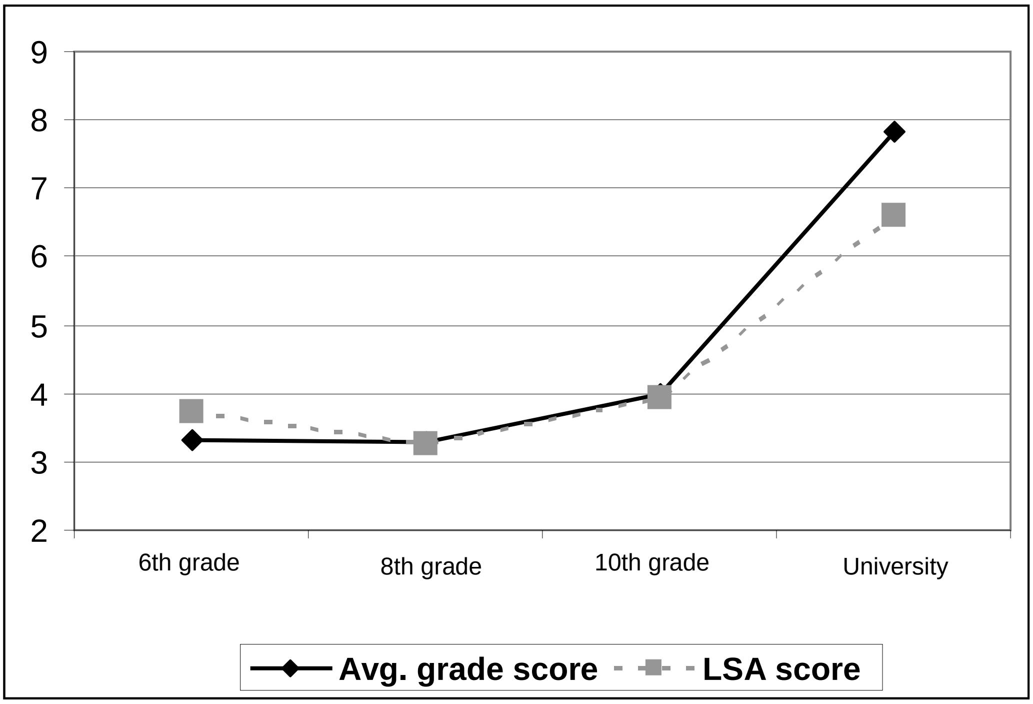

Figure 5. LSA and graders' mean score at each grade level for the expository text

An interesting phenomenon shown in figures 4 and 5 is that LSA scores overestimate the lowest grade level summaries and underestimate the university level summaries, while at the intermediate level it matches the human graders' opinions. This effect is clearest for the narrative text. LSA measures seem less appropriate for evaluating the lowest and highest scores than human graders

F) Correlations between multiple choice and summary assessments

Lastly, in this section our aim is to investigate the relationship between summary gradings and multiple choice scores. As shown by the data in Table 6, there are correlations between multiple choice scores and expert grader scores for summaries, quite similar for both texts. However, the correlations obtained for LSA are generally lower than those for experts. LSA scores using vector length vary greatly depending on the type of text, with no significant correlation for the narrative text and a significant positive correlation similar to those obtained with expert graders in the expository text.

	Human graders			LSA		
	Content	Coherence	Global (content + coherence)	Cosine	Vector length	Cosine + vector length
Narrative multiple choice test (N=249)	.37**	.45**	.43**	.18**	.11	.18**
Expository multiple choice test (N=243)	.36**	.41**	.39**	.18**	.37**	.34**

** p<.01.

Table 6. Correlations between multiple choice tests and summary assessments (judges and LSA) for each text and type of measure (cosine, vector length, content, coherence)

The results shown in Table 6 are homogeneous for human grader scores on both types of text and for both content and coherence, when compared with multiple choice scores. One conclusion that can be drawn from this data is that the graders scored in a very stable manner, independently of the text type. However, the correlation between multiple choice scores and LSA cosine and vector length measures were lower and more discrepant. The vector length measure seems more sensitive to the type of text studied here.

Conclusions

Summarizing is an extremely important task in facilitating students' comprehension of texts (Brown et al., 1983; E. Kintsch et al., 2007; van Dijk & Kintsch, 1983). The study presents a potential system for automated assessment of readers' summaries. This is an emergent approach to text analysis and automated grading that is gaining acceptance, as evidenced by intelligent tutors, automated grading systems, questionnaires, search systems, dialog management, etc.

Nonetheless, given the complexities of evaluating open-ended responses, it is still too early to declare the availability of a computer-based tool able to carry out these tasks automatically in an effective manner.

The aims of this study were threefold: (1) To obtain a reliable LSA-based method that combines vector length and the measure of semantic similarity (2) To compare the sensitivity of LSA and judges to differences in quality of summaries by different grade levels, type of text, and correlations with multiple choice test scores (3) To overcome the limitations of LSA with very short texts. The results showed progress toward fulfilling these aims, although in general they are relatively more satisfactory in evaluations of the expository than the narrative summaries studied in this paper.

On the one hand, we can conclude that the use of information on vector length together with semantic similarity provides an important improvement when evaluating highly conceptualized summaries with LSA. In previous studies where the texts graded were longer (Foltz et al., 1999; Landauer et al., 1998; Wade-Stein & E. Kintsch, 2004), high levels of reliability were achieved without the need for the measure of the vector length. But the limitations of LSA when applied to shorter texts (Rehder et al., 1998; Wiemer-Hastings et al., 1999) call for more information from the latent semantic space to be used. In this case, using semantic similarity with vector length we are better able to simulate evaluation by human graders. The additional information contributes to making LSA a more reliable tool. On the other hand, the results obtained show that with summaries of the expository text reliability is always above .70, and even above .80 when the average judges' scores are used. The results with the narrative text summaries are slightly less impressive, even though the reliability level was always above .60. It's common for the results to be weaker when working with narrative texts (Wolf, 2005), probably because they are less descriptive, less structured, less factual and more metaphorical than the expository texts. However, in an academic context it is more important to obtain reliable results with expository texts, which are prototypical academic texts. The better results for evaluations of the expository text translate into better sensitivity of LSA to differences in quality between summaries from different grade levels, more so than with the narrative text summary. Lastly, we should point out that a generalist corpus was used to obtain these results, the same as those on the University of Colorado official LSA webpage (http://lsa.colorado.edu). No *ad hoc* corpus was built to train LSA, and this should be encouraging news for researchers looking to apply and perhaps improve on this and other similar methods, without the need to incorporate their own subject area-specific corpus. We do not aim to generalize the results obtained to all types of narrative or expositive texts, but rather to generalize the regression method proposed for any kind of text to be evaluated with the LSA tool. We should note, however, that as indicated the success of the method is greater for expositive texts than for narrative texts. The method's improved performance on the expositive text could be partly due to it being slightly shorter than the narrative text, offering a certain advantage for the expositive text when we demand that the summaries are so concise.

One general finding from the results obtained is that LSA overestimates the lowest scores from lower grade levels, and underestimates the highest scores (university level), while for intermediate grade levels there is a closer match with the graders' criteria. This phenomenon is much clearer in the narrative text. One possible explanation is that previous knowledge is often greater in the narrative text, and there is greater usage of connectives and transitions that would exaggerate the differences between good and bad readers. Another factor may be the increased importance of implicit knowledge introduced by a greater number of inferences in narrative than expository texts (Graesser, 1981). On the other hand, these differences are perhaps lessened by the fact that in an expository text the summary more closely matches the content of the text, being less reliant on previous knowledge and with less usage of connectives, synonyms, etc. This should be borne in mind for future studies, to confirm whether the finding can be generalized.

This study also presents some limitations. The computer-based tool described in this paper is not yet ready for classroom implementation, so we must realistically consider what the current findings indicate for classroom prospects, and what still needs to be done to make this tool more useable in classroom applications. In addition, further research is required with more texts in order to form a sufficient basis for the generalization of our findings.

Nevertheless, the availability of automatic tools that evaluate reliably, help to detect weak or strong points in the summaries, may take pressure off of teachers, at the same time providing the student with assistance in everyday tasks. LSA has become one of the most widely-used computational tools of recent years, and one of the fastest-growing areas of application has been the field of education. Today, LSA is already a reality in some U.S. classrooms. It is gradually finding its way into schools to help improve students' writing and comprehension strategies (Dikli, 2006; E. Kintsch et al., 2007). This tool needs to be complemented with new algorithms to overcome some of its limitations (Kintsch, 2001; Jorge-Botana et al., 2009; Olmos et al., 2009), mathematically optimize the usage of the latent semantic space (Hu, Cai, Wiemer-Hastings, Graesser & McNamara, 2007). Also it is possible to joint several of them such as we implemented a combination between cosine and vector length. But also need it to link it with psychological models such as semantic memory (Denhière, Lemaire, Bellissens & Jhean-Larose, 2007; León, Jorge-Botana, Olmos & Escudero, 2010) or with other computational models of language (Steyvers & Griffiths, 2007). Once all of these contributions are added, the potential and the capability of LSA in the educational sector will be far greater.

ACKNOWLEDGEMENTS

This work was supported by Grant SEJ2006-09916 from the Spanish Ministry of Science and Technology and PSI 2009-31932 from the Spanish Ministry of Education.

References

Bellissens, C., Jeuniaux, P., Duran, N. D. & McNamara, D. S. (2010). A text relatedness and dependency computational model: Using Latent Semantic Analysis and Coh-Metrix to predict self-explanation quality. *Studia Informatica Universalis*, *8*, 85-125.

Bransford, J. D., Brown, A. L. & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. National Research Council Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Brown, A. L., Bransford, J. D., Ferrara, R. A. & Campione, J. C. (1983). Learning, remembering, and understanding. In J. Flavell & E. M. Markman (Eds.), *Handbook of child psychology* (4th ed.). *Cognitive development* (Vol. 3, pp. 515- 629). New York: Wiley.

Dessus, P. & Lemaire, B. (2002). Using production to assess learning: an ILE that Fosters Self-Regulated Learning. In S. A. Cerri, G. Gouardères & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems (ITS 2002)* (pp. 772-781). Berlin : Springer, LNCS 2363.

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.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5*, 1. Retrieved [date] from http://www.jtla.org.

Far, R., Pritchard, R. & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*, 209-226.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of*

Computer-Enhanced Learning, 1. Retrieve June 29, 2004, from http://knowledgetechnologies.com.

Graesser, A. C., Chipman, P., Haynes, B. C. & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education, 48,* 612-618.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H, Ventura, M., Olney, A. & Lowerse, M. M. (2004). Autotutor: A tutor with dialogue in natural language. *Behaviour Research Methods, Instruments, and Computers, 36,* 180-193.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z. & Hu, X. (2007) .Using LSA in AutoTutor: Learning through Mixed-initiative Dialogue in Natural Language. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.). *The Handbook of Latent Semantic Analysis* (pp. 243-262). Mahwah, NJ: Erlbaum.

Haley, D. (2009). Applying latent semantic analysis to computer assisted assessment in the Computer Science domain: a framework, a tool, and an evaluation. PhD thesis, The Open University.

Hu, X., Cai, Z., Wiemer-Hastings, Graesser, A. C. & McNamara, D.S. (2007). Strengths, limitations, and extensions of LSA. In T.K. Landauer, D. 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. (2009). Using LSA and the predication algorithm to improve extraction of meanings from a diagnostic corpus. *The Spanish Journal of Psychology, 12, 2*, 424-440.

Jorge-Botana, G., León J.A., Olmos, R. & Escudero, I. (2010). Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora. *Journal of Quantitative Linguistics, 17, 1*, 1-29.

Kakkonen, T. & Sutinen, E. (2011). EssayAid: Towards a Semi-automatic System for Assessing Student Texts. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*, 2-3,119-139.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N. & Dooley, S. (2007). Summary Street: Computer-Guided Summary Writing. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.). *The handbook of Latent Semantic Analysis* (pp. 263- 277). Mahwah, NJ: Erlbaum.

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.

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). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

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, Instruments and Computers 38*, 4, 616–627.

McNamara, D. S., Boonthum, C., Levinstein, I. B. & Millis, K. (2007). Evaluating Self-Explanations in iSTART: Comparing Word-Based and LSA Algorithms. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.). *The handbook of Latent Semantic Analysis* (pp. 227-242). Mahwah, NJ: Erlbaum.

McNamara, D. S., Levinstein, I. B. & Boonthum, C. (2004). iSTART: interactive strategy training for active reading and thinking. *Behaviour Research Methods, Instruments, and Computers*, *36*, 222-233.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S. & McNamara, D. S. (2007). Assessing and improving comprehension with Latent Semantic Analysis. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of Latent Semantic Analysis* (pp. 207-225). Mahwah, NJ: Erlbaum.

Monjurul Islam, Md. & Latiful Hoque, A. S. M. (2012). Automated Essay Scoring Using Generalized Latent Semantic Analysis. *Journal of Computers*, *7, 3*, 616-626.

Olmos, R., León, J. A., Jorge-Botana, G. & Escudero, I. (2009). New algorithms assessing short summaries in expository texts using Latent Semantic Analysis. *Behaviour Research Methods, Instruments, and Computers*, *41*, 944-950.

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

Quesada, J.F, Kintsch, W. & Gomez, E. (2001). A Computational Theory of Complex Problem Solving Using the Vector Space Model (part I): Latent Semantic Analysis, Through the Path of Thousands of Ants. In J. J. Cañas (Ed.) *Cognitive research with Microworlds,* (pp. 117-131). Granada (Spain).

Rehder, B., Schreiner, M. E., Wolfe, 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.

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

Steyvers, M. & Griffiths, T. (2007). Probabilistic Topic Models. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of Latent Semantic Analysis* (pp. 427-448). Mahwah, NJ: Erlbaum.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Wade-Stein, D. & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and 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.

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 and Cognition, 31*, 2, 359-364.

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