Article

Automated LSA Assessment of Summaries in Distance Education: Some Variables to Be Considered

Journal of Educational Computing Research
0(0) 1–24

DOI: 10.1177/0735633115571930
jec.sagepub.com

Guillermo Jorge-Botana[1], José M. Luzón[1], Isabel Gómez-Veiga[1], and Jesús I. Martín-Cordero[1]

Abstract

A latent semantic analysis-based automated summary assessment is described; this automated system is applied to a real learning from text task in a Distance Education context. We comment on the use of automated content, plagiarism, text coherence measures, and word weights average and their impact on predicting human judges summary scoring. A first regression analysis showed the independence of interparagraph coherence with respect to superficial text variables, advising its inclusion in a general regression model, along with content, plagiarism measures. The final regression model explains a considerable degree of variability in human judgment of summaries. Finally, we discuss several methodological implications and further applications of the automated summary scoring technique developed in this study.

Keywords

LSA, assessment, summaries, coherence, plagiarism, distance education

[1]Departamento de Psicología Evolutiva y de la Educación, Facultad de Psicología, Universidad Nacional de Educación a Distancia, Madrid, Spain

Corresponding Author:
Guillermo Jorge-Botana, Departamento de Psicología Evolutiva y de la Educación, Facultad de Psicología, Universidad Nacional de Educación a Distancia, C/Juan del Rosal, n° 10, 28040 Madrid, Spain.
Email: gdejorge@psi.uned.es

Introduction

In the last decade, latent semantic analysis (LSA) has been massively studied in its facet of the assessment of essays in the educational context (Bellissens, Jeuniaux, Duran, & McNamara, 2010; Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007; Jorge-Botana, León, Olmos, & Escudero, 2010; Olmos, León, Jorge-Botana, & Escudero, 2009; Rehder et al., 1998). The main application has been to evaluate free response questions in traditional and small-scale educational environments (directly in a student classroom or in a laboratory context; see Haley, 2009). Recently, some studies have used LSA in larger environments, integrated in the Learning Management System (LMS) of an institution (Foltz, Lochbaum, & Rosenstein, 2011). For this reason, it is important to study the technique in its own ecological context. Although there is extensive experience in the use of automated assessment tasks using multiple choice questions, or similar formats, in Distance Education the possibility of automated assessment of free response questions has been much less explored. This is somewhat of a paradox, considering the above-mentioned volume of research on the use of LSA in educational tasks. An explanation may be that there are some constraints to be considered when using automated assessment of free response questions in Distance Education: Students do not share a physical space with lecturers while working. For this reason, it is not possible to directly supervise students while they fulfill the task, and this fact increases the chances of less productive behavior patterns, mainly in the form of *copy-paste* answer strategies. In this sense, the detection of plagiarism becomes a crucial variable to be controlled and this increases systems demands. For instance, an ordinary procedure is the use of the shared n-grams between one essay and a reference text (Češka, 2010), that is, the proportion of bigrams or trigrams of words that is shared between two essays.

Evaluating With LSA

LSA is based on the concept of vector space models, an approach that uses linear algebra for allocating lexical units in an n-dimensional vector space. In general terms, LSA is a set of different processes by which a collection of texts (handbooks, reference texts, etc.), usually called corpus, is transformed into a semantic space. For educational purposes, it is more appropriate to use a corpus created from texts pertaining to the area of interest (for instance, Psychology if one is interested in assessing psychology student essays). This corpus is processed and then expressed into a matrix, which includes its terms and paragraphs (it is recommended to use paragraphs as text window; see Landauer, 2003; Rehder et al., 1998) in it. Then, a new step is applied in this matrix. A weight function, usually log-entropy, is carried on in order to prevent the great asymmetry in word frequency (Nakov, 2000; Nakov, Popova, & Mateev, 2001). In these calculations, a global entropy weight is assigned to each term.

A higher global weight tends to indicate a greater focalization. This index is also interesting because it gives some information about the terms. If a term lacks subject area focalization, its occurrence is not enough to predict the topic of our discourse. But what is most subtle about LSA and has made it famous is applying to this matrix a dimension reduction technique by means of singular value decomposition (SVD), which is a kind of eigenvectors and eigenvalues factorization, as for example, the Principal Component Analysis technique. This vector space is usually called *semantic space*. By means of this reduction, we can express each term or paragraph with only a few strong dimensions (dimensions which corresponding singular value are significantly big), and omit the others, avoiding what experts use to call ordinary language use noise (for details, see Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer, Foltz, & Laham, 1998). In this semantic space, some vector metrics as the cosine (or Euclidean distance) can be used to determine the semantic similarity of the terms and paragraphs (Landauer & Dumais, 1997). To represent new paragraphs, the simpler LSA method (see, e.g., Kintsch & Mangalath, 2011 for alternatives) rests on a single conceptually simple constraint, namely that the representation of any new meaningful passage must be composed as a function of the representation of the words it contains (Landauer, 2003, 2007). Thus, LSA models a passage as a simple linear equation composed with these term vectors. This technique is usually called folding-in method (see Martin & Berry, 2007 for details). This method is a very important piece in the automatic evaluation processes, as the student-constructed responses are usually projected in the semantic space by this way.

Student-constructed responses are those that require a student to produce a natural-language answer that may range from a couple of sentences to several paragraphs (Magliano & Graesser, 2012, p. 608). These responses (e.g., essays, summaries, short-answers, think-aloud protocols) provide a valuable opportunity to gauge student learning outcomes and comprehension strategies. Specifically, summarization has been extensively used both in educational practice to evaluate reading comprehension, as well as to activate reading strategies that promote deep comprehension. The summary task activates some processes that contribute to understanding a text: identify the significant text units, delete irrelevant details, select the main explicit ideas, and produce new ideas that activate the main ideas and integrate them with relevant prior knowledge and, finally, join them to produce a new short text, the summary (Brown & Day, 1983). It also illustrates individual differences in the extent to which readers enact them (Bereiter & Scardamalia, 1984; Graesser, Singer, & Trabasso, 1994; Kintsch, 1998). So summary elaboration as a cognitive task involves identifying main ideas pertaining to the situation to which the text refers.

Other relevant properties of summaries are the local and global coherence levels. In general, coherence is a function of the reader's ability to use various

strategies, such as making causal connections, drawing inferences, and applying background knowledge to assemble and link disparate pieces of information to each other and to prior knowledge (Lorch & O′Brien, 1995). Consequently, a connected mental representation is generally seen as arising from the coherence relations that are part of the reader's mental representation of the text (Mulder & Sanders, 2012; Sanders, Spooren, & Noordman, 1993). Coherence relations are defined as meaning relations (e.g., cause-consequence, claim-argument) between text segments (e.g., sentence-sentence, paragraph-paragraph). These semantic relations are often made explicit in the text by means of a linguistic marker (e.g., a connective); but, if this is not the case, the reader must infer the implicit coherence relations (Graesser, Millis, & Zwaan, 1997). Behavioral correlates have also been found. For instance, Wolfe, Magliano, and Larsen (2005) found a correlation between coherence measures and reading time. In addition, subjects with different amounts of knowledge benefit in different ways from different levels of coherence (e.g., McNamara, 1996). Therefore, coherence measures can be a central aspect in summary scoring. LSA has the capacity to model the quality of coherence and to quantify it by measuring the semantic similarity of one text section (e.g., sentence, paragraph) with respect to the next one. In fact, LSA is sensitive to proximal and distal semantic relationships. It has been applied to different types of discourse coherence examining both local coherence between sentences, and the global coherence of sentences to the essay as a whole (Foltz, Kintsch, & Landauer, 1998; Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007; McNamara, Cai, & Louwerse, 2007). In order to measure coherence by LSA means, individual text segments are represented as vectors in a semantic space and the cosine for the vector from each text section vector to the next one is computed. Generally, highly coherent discourse will have greater cosines, whereas less coherent discourse will have smaller cosines (Foltz, 2007). However, it may be noted that, as suggested by Foltz (2007), high levels of sentence-sentence coherence may receive low human scores.

Objectives

Our main purpose was to use and test a new LSA-based protocol to reliably evaluate written summaries produced by students in a Distance Education environment. The proposed protocol for computer-automated assessment was tested in the e-Learning Management System (e-LMS) of the National University of Distance Education of Spain (UNED). More precisely, the four central goals of this study were to:

1. Analyze whether the LSA-based protocol is sensitive enough to the quality of written summaries produced by students, as human graders are.
2. Examine how LSA can be applied to measuring different levels of discourse coherence (paragraph-paragraph; paragraph-sentence; sentence-sentence)

and to determine their efficiency in automatic evaluation of summaries to simulate human graders.
3. Provide an example of successful application of LSA in order to reliably evaluate real discourse tasks.
4. Estimate the use of a plagiarism measure as a component for the prediction of human judges' performance in summary assessment.

In line with prior findings explained above, the protocol incorporated different types of measures: (a) semantic content measures; (b) coherence measures (paragraph-paragraph; paragraph-sentence; sentence-sentence); (c) n-gram measures of plagiarism; (d) global weight measures for summaries. We predicted that there would be significant positive correlations between essential information derived from the latent semantic space and human graders' measures. Regression analysis would show that LSA predicts human scores, once the variables that better represent the human judgments of summary quality have been identified.

Method

Participants

The participants were 242 third-year undergraduate psychology students (24% males and 76% females) at the UNED. They volunteered in exchange for one hour's credit in the Educational Psychology course. Student age ranges were less than or equal to 25 years old, 25%; from 26 to 35, 21%; from 36 to 45, 32%; and over 45, 22%. In order to obtain human evaluations with which to compare those of LSA, two Psychology postgraduate students and one PhD in Psychology took part as judges and were paid for their involvement.

Materials

The reading text (the academic text that participant had to read and then summarize) was a fragment about reading acquisition from an Educational Psychology textbook (Santrock, 2012). It was 60 paragraphs, 3,660 words, and 296 lines long. It must be stressed that summarizing a text of this length is a plausible learning task in this context for this type of student.

The LSA model required the use of a large corpus. The corpus used in this study contains academic material sampled from various texts: In the field of Educational Psychology, we include the entire manual of the course (Santrock, 2012). In order to enrich the corpus and with the motivation that it will cover a more generic topics, we also include the Diagnostic and Statistical Manual of Mental Disorders-IV (1994), and some psychological texts from International Statistical Classification of Diseases and Related Health

Problems 10th Revision (CIE-10 in Spanish), and other expository texts in the field of Psychology. A semantic space with 300 dimensions, 16,317 words, and 11,566 paragraphs was used. To implement the evaluation protocol, a software called Gallito 2.0.12© (Jorge-Botana, Olmos, & Barroso, 2013) was used, which allowed us to compute a semantic space, evaluate the summaries, and measure the LSA and n-gram variables.

Procedure

The text to be read, that in this study is denominated as reading text, was available to each student as a Word text file via the online e-LMS in use at UNED (aLF). The task required students to read the text and summarize it (with no time limit in this part of the task). Next, students were asked to return to the e-LMS and post their summary. The instructions allowed students to use their spontaneous strategies to produce the response (e.g., a personal elaboration, a copy-paste strategy to summarize the original text, etc.). The size of the summary was not limited.

Human evaluations of summaries were collected from three judges. They were asked to take several features into account as evaluation criteria: the quality of the content (i.e., main ideas), use of adequate writing style, length, use of correct technical words, coherence, and personal elaboration of the summary produced (i.e., lack of very close paraphrases of the textual sentences that were just read). Each summary was graded independently by each of the three judges on a scale of 0 to 10. The judges carried out the evaluation independently and without knowing the writer's identity. An average score was also obtained from the three graders' scores for each summary.

Protocol for automated essay scoring. To implement and test the automated assessment protocol, we used a database comprising 242 summaries. The procedure to assess the proposed LSA-based and n-gram protocol incorporates several fundamental steps.

First, the protocol attempts to provide an assessment of plagiarism with n-gram measures. Second, LSA content semantic measures were obtained to assess the level of similarity to the source text. Students' summaries reflect a broad range of responses that vary in their content similarity with an expert summary, usually called *golden summary* (Landauer et al., 1998). In addition, coherence and surface features (e.g., number of words per each predefined unit such as sentence, paragraph, or summary) for each answer were also scored and analyzed. A global weight measure for each content word was also obtained to assess a global weigh ratio for every summary. Finally, once reliable evaluations combining information derived from all measures were computed, a regression equation based on them aimed at predicting human judges' scores for the summaries was applied and validated.

Plagiarism measures. For our purposes, it is interesting to evaluate how effectively the summary covers content from the text and whether the reader is appropriately sourcing where its ideas come from or whether he or she is plagiarizing the text. Of course, plagiarism is not an all/none process, but rather it is better construed as a continuum that ranges from extreme originality, or idiosyncrasy, to the literal transcription of the source text. So one of the challenges is to determine to what degree a text should be considered plagiarism from an academic point of view.

A useful approach is to use algorithms that consider sequences of n-grams and structured configurations of words (Cai et al., 2004). These algorithms can compute overlap of n-grams (bigrams for pair words, trigrams for word triplets, and so on) in their assessment of text overlap. In this study, plagiarism scores are based on the proportion of matches between trigrams in the reading text and the student's summary in order to detect plagiarism; the use of trigrams is supported by previous research (Barón-Cedeño & Rosso, 2009). Trigrams were computed by using the Teraman.Net (Češka, 2010; Češka, Hanák, & Tesař, 2007), a library for n-gram extraction available for the .NET Framework, which was integrated with the main software used to calculate the measures.

Two kinds of measures were used to operationalize plagiarism: (a) *Recall,* that is, the proportion of original source text that it is present in each summary (overlap of trigrams); (b) *Precision,* that is, the proportion of summary that it is present in the original source text (overlap of trigrams). These two measures-recall and precision-ranged from 0 to 1 and were combined in an index, F'. These Recall and Precision scores were then transformed by calculating a new score for both variables: 1 minus *Precision* (ρ_i in equation (1)) and 1 minus *Recall* (π_i in equation (2)). They were combined in an index F' (F_i in equation (3)), where *Original* is the original source text, $Summary_i$ is a student's summary. Consequently, the higher the new transformed score was, the more elaborated (i.e., more original) or the less plagiarizing was the summary.

$$\rho_i = 1 - \frac{\left|Original \cap Summary_i\right|}{\left|Original\right|} \tag{1}$$

$$\pi_i = 1 - \frac{\left|Original \cap Summary_i\right|}{\left|Summary_i\right|} \tag{2}$$

$$F_i = \frac{2\pi_i \rho_i}{\pi_{i+}\, \rho_i} \tag{3}$$

This kind of index is frequently used to measure the degree of success in Internet search-engines for information recovery. They are also used for tasks similar to the one used in this study, for instance, in the matching of

spontaneous speech recognition (Jorge-Botana, Olmos, & Barroso, in press; McCowan et al., 2004).

LSA content measures. LSA measures were obtained for the assessment of summaries. Once the space was created, as was outlined in the Introduction section, the semantic content of summaries was evaluated. For this purpose, we used a well-known method usually applied in Automated Essay Scoring with LSA: the *expert method* (see, e.g., Foltz, Laham, & Landauer, 1999; Kintsch et al., 2007; Landauer & Dumais, 1997, 2008; León, Olmos, Escudero, Cañas, & Salmerón, 2006; Olmos et al., 2009). This method consists in assessing student's summaries by comparing them with a summary written by an expert, usually known as the golden summary. It is conceived as a method to capture the semantic similarity between each student's summary and the golden summary (see Rehder et al, 1998).[1] More precisely, LSA computes the similarity between the vector that represents each summary and the vector that represents the golden summary. Both summaries are projected on the semantic space by means of folding-in method. In this study, an expert (a PhD in Educational Psychology) read the reading text and wrote the golden summary. This golden summary was evaluated by two experts, who also hold PhD degrees in Educational Psychology, and they both agreed that it was an ideal summary that contained all the relevant information in the reading text and had a high degree of coherence.

It should be noted that the semantic similarity between the golden summary and the student's summaries was scored as the *Euclidean distance* between the vectors associated. This distance measure was selected because it represents the semantic similarity and is also affected by the vector *length* component associated with each summary. The more semantically similar the student summary to the golden summary, the higher the LSA-Content score. Prior studies have evidenced the higher efficiency of the Euclidean distance measure compared with cosines between vectors in automated evaluation of summaries (see Jorge-Botana et al., 2010; Olmos et al., 2009).

LSA coherence measures. An LSA-based measure determines coherence using the derived semantic similarity of one text unit to the next. This is performed both between paragraphs and between sentences (Foltz, 2007). Following Foltz et al. (1998), taking the mean of the cosines of LSA vectors representing successive units in a text, it is possible represent coherence. In LSA, highly coherent discourse should be represented by greater cosines, whereas less coherent discourse should have smaller cosines.

First, paragraph-paragraph coherence was computed as the cosine between each paragraph and the next. Then, the average cosine between any two adjacent paragraphs for each summary was computed (equation (4)). Paragraphs were defined by carriage returns. Only paragraphs 10 or more words long were considered for data analysis, so titles and notes were excluded from analysis.

Second, sentence-sentence coherence for each paragraph was computed as the cosine between each sentence and the next, in order to provide a local measure of coherence within a paragraph (equation (5)). So sentence-sentence coherence was assessed as an intraparagraph measure and did not include cosines between sentences of different paragraphs; paragraphs with just one sentence were not considered for analysis either. The average sentence-sentence coherence for all paragraphs in the summary was calculated (equation (6)).

Third, sentence-paragraph coherence was computed for each sentence within a paragraph.[2] This third coherence measure identifies the similarity between each sentence and the paragraph that contains it. Again, the sentence-to-paragraph coherence for each paragraph was computed (equation (7)) and an average coherence for all paragraphs in the summary (equation (8)). There are some caveats to be considered about this latter coherence measure because it could be too dependent on the number of sentences in the paragraph. To check the dependency of each coherence measure on the surface variables of the text, we computed the following data for each summary: number of paragraphs, number of words, mean of words per paragraph, mean of words per sentence, and mean of sentences per paragraph. We assume that, given the abstract nature of the construct of coherence, its optimal measure should be the one that proves to be most independent from surface variables in the text. To this end, we computed regression equations in which the dependent variable was each one of the coherence measures and the predictors were each of the surface variables. More on this later.

$$Cohp = \frac{\sum_{i=1}^{n-1} \cos(P_i, P_{i+1})}{n-1}. \tag{4}$$

where *Cohp* is the paragraph-paragraph coherence, P_i and P_{i+1} are each paragraph and the next one in the summary, and n is the number of paragraphs.

$$Cohss_j = \frac{\sum_{i=1}^{n-1} \cos(S_i, S_{i+1})}{n-1} \tag{5}$$

$$Cohss_{total} = \frac{\sum_{j=1}^{p} Cohs_j}{p} \tag{6}$$

where $Cohss_j$ is the sentence-sentence coherence for paragraph j, S_i and S_{i+1} are each sentence and the next in the paragraph j, n is the number of sentences in the paragraph j; and $Cohss_{total}$ is the average sentence-sentence coherence for a summary, and p is the total number of paragraphs in the summary.

$$Cohsp_j = \frac{\sum_{i=1}^{n} \cos(S_i, P_j)}{n} \tag{7}$$

$$Cohsp_{total} = \frac{\sum_{j=1}^{p} Cohsp_j}{p} \tag{8}$$

where $Cohsp_j$ is the sentence-paragraph coherence for paragraph j, S_i and P_j are each sentence i and the paragraph j that contains it, n is the number of sentences in paragraph j, and p is the total number of paragraphs in the summary.

Global weight ratio. As was explained above, LSA uses a transformation of the raw frequency of word-context co-ocurrence (see, e.g., Landauer & Dumais, 1997) that is called log-entropy. Its goal is to decrease the influence of words that are extremely frequent and that do not carry much meaning. The measure used in this study to assess how informative the language used by students in their summaries was the *global weight ratio* for each word. The global weight gives more importance to words that convey more information within the corpus and is part of the common log-entropy formula (equation (9)). The global weight ratio for each summary (equation (10)) is a way to assess the informativeness and the synthesis competence expressed in each summary. In other words, the mean global weight is used to assess the use of an appropriate technical lexicon and the lack of trivial examples.

$$\begin{aligned} g_i &= 1 + \sum_j \left(\frac{p_{ij} \log(p_{ij})}{\log n} \right) \\ p_{ij} &= \frac{tf_{ij}}{gf_i} \end{aligned} \tag{9}$$

where g_i is the global weight for the term i, and tf_{ij} is the number of occurrences of term i in the corpus document j, gf_{i} is the total number of times term i occurs in all corpus documents, and n is the number of documents in the corpus.

$$g_{total} = \frac{\sum_{i=1}^{w} g_i}{w} \tag{10}$$

where g_{total} is the global weight for a summary, and W is the number of words in a summary.

Results

Data were analyzed in several stages. First, the correlations between the human judges' scores were obtained. Likewise, the correlations between LSA and human scores for the summary task were calculated. Next, a regression equation was obtained for the coherence measures and surface text variables. Then, a

regression equation was obtained using a training sample. Finally, the model was validated using the rest of our sample of summaries.

Correlations Between Human Judges' Scores

For each student summary, three independent human judges rated its quality, and a mean quality rating across judges was computed for each summary. Before analyzing the validity of LSA-based measures, correlations among the three judges were analyzed. As expected, there was a clear pattern of positive correlations between judges (J1–J2: $r = .47$, $p < .001$; J1–J3: $r = .49$, $p < .001$; and J2–J3: $r = .38$, $p < .001$, two tailed in all cases). Prior studies (see, e.g., Graesser et al., 2007) found somewhat higher correlations between human raters. A plausible explanation for our lower interjudge correlations in this study has to do with the human judge's lack of sensibility to plagiarism (more on this ahead). Besides, it should also be considered the kind of evaluation criteria used. Remember that human judges were asked to consider in their global score features such as the quality of the content, use of adequate writing style, length, technical words, coherence, and elaboration, but in order to provide ecological validity, we did not provide them with a rubric and specific grading criteria for these features.

As can be observed from Table 1, the use of different plagiarism thresholds shows a remarkable increase of interjudge correlations. Human judge's low sensibility to plagiarism may again be a plausible explanation for these results. These threshold values were empirically determined from the plagiarism mean least ***i*** times ¼ Standard deviation ($i = 0, 1, 2, 3, 4$, respectively). In our opinion, these results suggest that the human judges were not equally sensitive to plagiarism detection.

Correlations Between Human Judges' Mean Scores and Automatic Measures

Due to such interjudge medium agreement aimed in the last section, we took a first methodological decision: to calculate correlations between human scores

Table 1. Reliability (Pearson Correlation) Between Judges for Different Threshold Plagiarism Values: Mean (.50) Least *i* Times ¼ Standard Deviation (.27).

	$i = 0$	$i = ¼$	$i = ½$	$i = ¾$	$i = 1$
Plagiarism threshold	**.50**	**.43**	**.36**	**.29**	**.22**
Judges 1–2	.47**	.49**	.53**	.57**	.57**
Judges 1–3	.58**	.60**	.64**	.68**	.72**
Judges 2–3	.36**	.37**	.42**	.46**	.51**

Note. **Correlation is significant at the .01 level (two-tailed).

Table 2. Reliability (Pearson Correlation) Between Human Judges' Mean Quality Ratings (Human), LSA-Content, LSA-Coherence (Sentence-Sentence, Sen-Sen; Sentence-Paragraph, Sen-Par; Paragraph-Paragraph, Par-Par), Plagiarism, and Global Weight.

$n = 208$	Human ratings	LSA-Content	Sen-Sen	Sen-Par	Par-Par	Plagiarism	Global weight
Human ratings	1						
LSA-Content	.59**	1					
Sen-sentence	−.22**	−.32**	1				
Sen–paragraph	−.25**	−.22**	.56**	1			
Par-paragraph	.09	−.10	.44**	.12			
Plagiarism	−.24**	−.25**	.33**	.18*	.19**	1	
Global weight	−.17*	−.05	.22**	.13	.19**	.07	1
M	6.41	2.94	0.34	0.35	0.44	0.506	0.47
SD	1.02	0.66	0.08	0.04	0.07	0.27	0.01

Note. *Correlation is significant at the .05 level (two-tailed). **Correlation is significant at the .01 level (two-tailed).

and automatic measures, and also the regression equation to predict human scores, only with summaries that presented a sufficiently high interjudge agreement. For this purpose, we estimated a new variable which measures Inter-Judge dispersion: the standard deviation between the three scores of the judges. We called it JDispersion ($M = 1.50$; $SD = 0.620$). Our exclusion criterion was to consider outlier any summary which presented 1 *SD* above the mean of the JDispersion distribution. Thus, from an initial sample of 242 summaries, we worked after outlier elimination on a final sample of 208 ($M = 1.329$; $SD = .456$).

With this new summary selection, we calculated the correlations between automatic scores and the mean of the judges' quality ratings (Table 2). As predicted, LSA-Content reliably correlated with the human judges' score (.59). Prior studies also evidenced that correlations between the LSA scores and the judges' quality ratings were approximately $r = .50$ for expository texts (e.g., Graesser et al., 2000; Olde, Franceschetti, Karnavat, Graesser, & the Tutoring Research Group, 2002; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999); although other studies have yielded higher correlations (Haley et al. 2009). It should be noted that the correlations between one pair of judges in our study were approximately .50 (see Table 1); LSA-Content agrees with the judges approximately as much as any two judges agree with each other.

The correlation between the mean of the judges' quality ratings (Human) and sentence-paragraph (Sen-Par) was reliably negative, indicating that the higher intraparagraph coherence, the lower the human score. Also, the correlation between the human ratings and sentence-sentence (Sen-Sen) was negative.

Table 3. Descriptive Statistics.

Surface variables	*M*	*SD*
Number of words	446.92	258.36
Number of paragraphs	18.51	10.33
Number of sentences per paragraph	2.17	1.02
Number of words per paragraph	26.87	19.14
Number of words per sentence	13.47	3.35

It is assumed for some authors that LSA coherence measures seem to go contrary to humans scores (Bestgen, Lories, & Thewissen, 2010), but in our case, we want to highlight that paragraph-paragraph coherence has no negative correlation with human ratings and seems to behave in a different way as the other coherence measures.

Also, the correlation between paragraph-paragraph and sentence-sentence shows that intraparagraph coherence (the sentence similarity inside a paragraph) perhaps has the effect of enhancing interparagraph coherence.

On other hand, it should be noted that the correlation between paragraph-paragraph coherence and the global weight ratio was also reliably positive, indicating that the higher the coherence between paragraphs, the higher the word weight in summaries. Correlations between coherence and plagiarism were reliably positive for the coherence measures.

The observed correlation between global weight ratio and coherence may result from student's keeping to the point and not making an excessive use of examples, which perhaps contributes to a more coherent text. Some studies have also found similar effects of lexical diversity and interpreted it as an artifact because the mere repetition of words in sentences increases coherence as measured by LSA (Bestgen et al., 2010). In our study, global weight ratio does not imply the repetition of words, but the use of a high proportion of informative words, which means that this variable may not be an artifact but represent an adequate strategy for keeping the fluency of the text. However, global weight ratio seems to show a negative correlation with human scores, which makes it difficult to interpret. Perhaps human judges tend to give inferior scores to texts that appear too synthetic. Nevertheless, we offer both sets of data and leave it to further research a more precise interpretation of their meaning.

Regression Analyses for Coherence

In order to assess the capacity of the surface properties of the text measures to predict coherence measures (Table 3), we carried out three multiple regression analyses on the main variables: coherence sentence-sentence, coherence sentence-paragraph, and coherence paragraph-paragraph (Table 4). Considering

Table 4. Results of the Regression Analyses for the Significant Variables.

Dependent variable: *Sentence–sentence coherence*			
Significant model $F(5, 228; 9.974; p < .0001)$; explained variance 16.1% (Adjusted $R^2 = .161$)			
Variables model	*B*	*SE B*	β
Number of words	7.863E-5	.000	.251
Number of paragraphs	1.850E-5	.001	.002
Number of sentences per paragraph	.010	.011	.128
Number of words per paragraph	.000	.001	.059
Number of words per sentence	.007	.002	.276**
Dependent variable: *Sentence–paragraph coherence*			
Significant model $F(5, 228; 27.320; p < .0001)$; explained variance 36.1% (Adjusted $R^2 = .361$)			
Variables model	*B*	*SE B*	β
Number of words	−2.151E-5	.000	−.140
Number of paragraphs	.001	.001	.295*
Number of sentences per paragraph	−.030	.005	−.777**
Number of words per paragraph	.001	.000	.353**
Number of words per sentence	.001	.001	.078
Dependent variable: *Paragraph–paragraph coherence*			
Significant model $F(5, 228; 12.453; p < .0001)$; explained variance 19.7% (Adjusted $R^2 = .197$)			
Variables model	*B*	*SE B*	β
Number of words	3.504E-5	.000	.136
Number of paragraphs	−9.249E-5	.001	−.014
Number of sentences per paragraph	.002	.026	.017
Number of words per paragraph	.003	.002	.398

*$p < .05$. **$p < .001$.

that this is an automatically assessed variable, present in the summaries with independence of the judges' scores, we have used the full sample of summaries. The predictive variables were number of words and number of paragraphs per summary, mean number of words per sentence and per paragraph, and mean number of sentences per paragraph (Table 3).

The results indicate that only paragraph-paragraph coherence was independent from the surface variables in the summary, whereas sentence-sentence coherence and sentence-paragraph coherence were dependent on surface variables.

Table 5. Results of the Regression Analyses for the Significant Variables.

Dependent variable: *Human judges' score*				
Significant model $F(3, 164; 32.013; p < .0001$; explained variance 35.8% (Adjusted $R^2 = .358)$				
Variables model	*B*	*SE B*	β	Tolerance
(Constant)	7.449	.480		
LSA-Content	.743	.086	.552**	.94
Par-paragraph coherence	3.288	1.017	.208**	.93
Plagiarism	−.594	.249	−.157*	.98

*$p < .05$. **$p < .001$.

Based on these findings, we introduced only the paragraph-paragraph coherence in the following regression model.

Regression Model Training

A regression analysis (least squares) was performed to evaluate the capacity of LSA-based measures to predict judges' scores. Because our target is predicting human scores, we used the more reliable sample of summaries with no outliers ($n = 208$, remember that we considered outlier any summary that presented 1 *SD* above the mean of the JDispersion distribution). From this sample, a random training sample (80%) was used to find a regression equation that was later applied to predict judges' scores and validate the proposed model with the 20% rest of summaries (the validation sample).

LSA-Content (LSA-Con) (i. e., semantic similarity), paragraph–paragraph LSA-Coherence (Par-Par), and plagiarism (P), while the average human judges' score was the dependent variable. Global weight ratio (global weight) can be considered in our study an exploratory variable, which implies a more recent and more theoretically debatable construct. Besides, perhaps because of this and as shown by its correlation matrix with the rest of our variables, the interpretation of its empirical relationships is not yet too clear. Thus, after having tested a first model, which included this variable, and proved a poorer predictive value, we did not include global weight ratio (global weight) in our second regression model, which is the one we present below. Table 5 shows the results of the regression analyses for the significant variables.

Regression Model Validation

To predict the average human judges' scores, the remaining sample of summaries was used to validate the model (validation sample, 20%). The automated scores of LSA-Content (LSA-Con), LSA-Coherence paragraph-paragraph

(Par-Par), and plagiarism (P) were obtained for each summary. Next, the above proposed regression model was applied to predict human scores. The correlation between the human judges' ratings and the scores predicted by the model was reliable ($r = .72$, $p < .001$); 52% of the variance in human scores was predicted by the model. It is clear that when the judges give the summary a low score, our regression model does the same; our model also follows suit when the judges give a high grade. It seems that there were no summaries whose predicted score grade differed markedly from the human grade. It should be remarked that the predictive power of LSA-Content alone for this sample was $r = .67$ ($p < .001$), so the regression model (LSA-Content together with LSA-Coherence paragraph-paragraph, and plagiarism) enriched the assessment.

Discussion

Evaluation of student writing often involves assessing both how well a topic is covered by a student and how coherently the student has linked the ideas within the summary. This study presents a potential protocol that could be used to automatically assess written summaries of expository texts produced by students in a Distance Education environment. In this regard, we would like to make several main points.

First, we analyzed whether the proposed protocol is sufficiently sensitive to the quality of written student responses, as human graders are. The results indicate that our protocol can be successfully used to evaluate summaries and identify a quality threshold value that is good enough to classify student responses as either being of low or medium-high quality. Moreover, the LSA-based protocol can be used to simulate the summary quality ratings of human judges. The results indicate that LSA measures such as semantic similarity (distance between the golden summary and the student summary) and paragraph-paragraph coherence, as well as plagiarism F' index, predicted human judges' scores when submitted to regression analyses. Our final regression analysis included the following variables: LSA-Content, LSA Paragraph-Paragraph coherence, and plagiarism. The results showed that these three measures explain a sizeable part of the total variance (52%) of human judges scoring. Therefore, the proposed model takes us a step closer to a system based on these variables that automatically evaluates summaries, offering us a close resemblance to human judges' performance.

As expected, the correlation between LSA content measures and human judges' scores was moderate but significant, indicating that the lower the distance between the golden summary and the student response, the higher the score given by human judges. It may be noted that LSA does not measure plagiarism per se, so LSA would find that a partially plagiarized summary approximately includes the original text main content. This is the reason why we think it is so important to fit a regression equation to include more variables,

specifically a plagiarism index, and not just the proximity between LSA and the golden summary. Moreover, it is also advisable to establish a plagiarism filter on to select which summaries will be regarded as adequate for analysis, as LSA would be *overrating* a summary that the human judges regarded as too literal. With respect to the moderate correlation between the human judges' scores, it may be that humans are not particularly sensitive to plagiarism, or not especially apt for its efficient detection. Along this line of thought, previous studies have reported human biases when asked to assess a product, an activity, or a text as novel (Defeldre, 2002). These studies apply the concept of *criptomnesia* to the process of regarding a text as novel when in fact it is not. This process has also been described as involuntary or unconscious plagiarism. Our results show that when the plagiarism threshold is suitably raised to identify summaries that should be excluded from analysis, interjudge correlation rises dramatically.

Second, a valuable contribution of this study is that LSA measures of coherence are incorporated into the regression model to predict human scores. We aimed to examine how LSA can be applied to measuring different levels of discourse coherence (paragraph-paragraph; paragraph-sentence; sentence-sentence) and to determine their efficiency in automatic evaluation of summaries to simulate human graders. As it was shown, LSA can be applied to the measurement of various levels of discourse coherence, and our findings established that paragraph-paragraph coherence is a particularly efficient variable in automatic evaluation of summaries to simulate human graders. It is relatively independent from surface measures and explains some variance in the regression function.

However, when we consider a different unit of analysis such as sentence-sentence coherence, the results indicate that the lower sentence-sentence coherence is, the higher the score assigned by human judges, a result that is congruent with the findings of other studies displaying the same results using sentence-sentence coherence (Bestgen et al., 2010). In fact, Foltz (2007) has warned that sentence-sentence coherence sometimes does not seem to match the scores assigned by human judges. However, contrary to the interpretation given in Bestgen et al. (2010), we take this result to entail that human judges do not consider local levels of coherence, but rather have a more global and semantic concept of the transitions in the text.

Another aspect that was suggested by paragraph-paragraph coherence is that we assumed that the best measure of coherence would be the one least dependent on the surface properties of the text. This is more obvious if we regard coherence as an emergent property of the text that transcends surface traits. To establish this, we estimated three regression equations, using each of the three measures of coherence as a dependent variable, and each of the surface properties of the text as a predictive variable. Our results confirm that paragraph-paragraph coherence is independent from surface variables, making it more reliable and allowing flexible use with different types of texts and styles. In summary, paragraph-paragraph

coherence makes more theoretical sense, is backed by studies that positively relate it to human judges' scores, and seems independent from the surface variables of the text. For all these reasons, it was a good candidate to be included in the regression model. Furthermore, it showed a positive correlation with Global Weight Ratio, suggesting that the more relevant information appear in the paragraph, the higher paragraph-paragraph coherence is. In fact, this very same Global Weight Ratios has been used to optimize measures of coherence (McNamara, Boonthum, Levinstein, & Millis 2007).

Third, the results give support to the assumption that LSA-based measures allow educationally relevant student summaries assessment. For the purpose of our research, we differentiated between two types of systems that fulfill different functions in reading assessment, depending on whether responses reflect meaning making (e.g., self-explanations) or products of comprehension (e.g., summaries). The main goal of systems developed to assess meaning making is to focus on students' active generation of information; by contrast, the purpose of systems developed to provide insights into the quality and nature of students' comprehension levels is to achieve reliable or valid information. Our research focuses on the latter type; specifically, on the grading of expository text summaries as a result of a reading comprehension task. Although developing automated scoring for this kind of student responses is a challenge, prior studies evidenced that some automated systems are as reliable as human raters (e.g., Jorge-Botana et al. 2010; Landauer, 2002; Landauer et al., 2003; Shermis, Burstein, Higgins, & Zechner, 2010).

To provide an example of successful application of LSA in order to reliably evaluate real discourse tasks, we propose a set of measures that are easily and automatically calculated and can be interpreted as revealing deep comprehension. Thus, a summarization task together with its LSA-based analysis could provide an alternative to standard measures such as multiple-choice tests for assessing some aspects of reading comprehension. LSA can grade the protocols (summaries) level of understanding by comparing with an expert summary representing a deep level of understanding. Using LSA to grade student responses in relation to the main ideas of the text and its structured expression in the summary is advantageous because summaries assessment is time consuming for human judges in educational environments with a high number of students.

However, we must say that LSA implies some limitations for tasks intended to assess reasoning processes or summaries which call for elaborative inferences as a relevant aspect (Kurby et al, 2003; Millis et al., 2004; Wolfe & Goldman, 2003). Nevertheless, LSA can be useful as part of a learning object or as part of a training and assessment process, given that the student's activity is explicitly oriented and referred to a well-defined content. For instance, LSA can be applied to situations in which we do not intend a comprehensive and formal grading but a formative evaluation on some relevant competence such as summarizing a text. This could allow for a much more frequent use of writing tasks

not only as an assessment tool but also as a learning activity, especially in contexts such as distance education, in which the student/teacher ratio usually precludes frequent essay writing. In fact, some authors have stressed that a text can be considered as understood only if the student is able to adequately summarize it (Palincsar & Brown, 1984). A good example of this philosophy of writing is Summary Street (Wade-Stein & Kintsch, 2004) or MyWritingLab (www.mywritinglab.com/learn-about/what-is.html). In both cases, the student is prompted to summarize a certain topic, but summaries are not marked by a human judge; it is the automated assessment of the texts that offers a feedback on the content and its structure. These tools have also shown a powerful motivational effect, producing high levels of student participation and satisfaction (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; Kintsch et al., 2007).

Finally, another valuable contribution of this study is that it provides an example of successful application of LSA in the evaluation of real tasks in a distance learning environment. A potential practical implication is the availability of automatic tools that reliably evaluate, helping to detect weak and strong points in summaries, offering new activity opportunities for teachers and students while, at the same time, providing students with individual feedback and tutoring. This view has been gaining acceptance in recent years (see, e.g., Foltz et al., 1999; Jorge-Botana et al., 2010; Kakkonen & Sutinen, 2011; Landauer, 2003; McNamara, Levinstein, & Boonthum, 2004; Monjurul & Latiful, 2012; Srihari et al., 2008).

In sum, our results confirm the successful application of LSA in the evaluation of real tasks in a distance learning environment, as well as the capacity of LSA-based measures in predicting human scores of written summaries produced by students. It also underscores the role of coherence in general, and paragraph-paragraph coherence in particular in the development of an automated protocol to assess free response questions. Likewise, our results highlight the importance of analyzing more variables, specifically a plagiarism index, and not just the proximity between LSA and the golden essay. However, in order to draw some more precise conclusions that allow for various improvements in computational techniques applied to education and cognitive sciences, wider experimental work with students across a diversity of tasks is clearly necessary. Finally, our focus on LSA-automated assessment reveals our fundamental interest in how computers actively assist the human cognitive resources while participating in learning activities.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Other procedures use more than one golden summary, so that eventually the mean of all similarities is computed for each of the summaries produced by the participants. The comparison between the participant's summary and the golden summaries makes it possible to assess the quality of the participant's answer and which relevant content is absent from it (see León et al., 2006).
2. This measure was also used to identify the most representative sentence within a paragraph (Kintsch, 2002).

References

Barrón-Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. In Boughanem, et al. (Eds), *Advances in Information Retrieval, ECIR 2009, LNCS 5478* (pp. 696–700). Berlin/Heidelberg, Germany: Springer-Verlag.

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.

Bereiter, C., & Scardamalia, M. (1984). Information processing demand of text composition. In H. Mandl, N. Stein & C. Trabasso (Eds), *Learning and comprehension of text* (pp. 407–428). Hillsdale, NJ: LEA.

Bestgen, Y., Lories, G., & Thewissen, J. (2010). Using latent semantic analysis to measure coherence in essays by foreign language learners? In S. Bolasco, I. Chiari & L. Giuliano (Eds), *Proceedings of 10th international conferences journée d'analyse statistique des données textuelles*, (JADT2010), 9–11 July 2010, Rome.

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

Cai, Z., McNamara, D. S., Louwerse, M. M., Hu, X., Rowe, M. P., & Graesser, A. C. (2004). NLS: A non-latent similarity algorithm. In K. D. Forbus, D. Gentner & T. Regier (Eds), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 180–185). Mahwah, NJ: LEA.

Češka, Z. (2010). *Automatic plagiarism detection based on latent semantic analysis: Theory and practice* (pp. 101–110). Saarbrüecken, Germany: VDM Verlag Dr. Müller.

Češka, Z., Hanák, I., & Tesař, R. (2007). *Teraman: A tool for N-gram extraction from large datasets. ICCP 2007* (pp. 209–216). New York, NY: IEEE.

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.

Defeldre, A. C. (2005). Inadvertent plagiarism in everyday life. *Applied Cognitive Psychology*, *19*, 1033–1040.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. Landauer, D. McNamara, D. Simon & W. Kintsch (Eds), *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*, *1999*(1):939–944.

Foltz, P. W., Lochbaum, K. E., & Rosenstein, M. B. (2011). *Analysis of student writing for a large scale implementation of formative assessment*. Paper presented at the National Council for Measurement in Education, New Orleans, LA.

Graesser, A. C., Millis, K. N., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*, 163–189.

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.

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

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & The Tutoring Research Group. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 129–148.

Haley, D. (2009). Applying latent semantic analysis to computer assisted assessment in the computer science domain: A framework, a tool, and an evaluation (Doctoral dissertation). Buckinghamshire, UK: The Open University.

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., 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.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013, July). Gallito 2.0: A natural language processing tool to support research on discourse. *Proceedings of the 13th Annual Meeting of the Society for Text and Discourse*. Valencia, Spain.

Jorge-Botana, G., Olmos, R., & Barroso, A. (in press). Call routing based on a combination of the construction-integration model and latent semantic analysis: A full system. *Informatica*.

Kakkonen, T., & Sutinen, E. (2011). Essay aid: 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. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

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, the Netherlands: Benjamins.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, *3*, 346–370.

Kurby, C. A., Wiemer–Hastings, K., Ganduri, N., Magliano, J. P., Millis, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent

semantic analysis space for science text. *Behavior Research Methods, Instruments and Computers*, *35*, 244–250.

Landauer, T. K. (2002). On the computational basis of cognition: Arguments from LSA. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 43–84). New York, NY: Academic Press.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education*, *10*(3):295–308.

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds), *Handbook of latent semantic analysis* (pp. 3–34). New York, NY and London, England: Routledge-Taylor & Francis Group.

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., & Dumais, S. T. (2008). Latent semantic analysis. *Scholarpedia*, *3*(11):4356.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning human-like knowledge by singular value decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds), *Advances in neural information processing systems* (Vol. 10, pp. 45–51). Cambridge, MA: MIT Press.

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.

Lorch, R. F. & O'Brien, E. J. (Eds),. (1995). *Sources of coherence in reading*. Hillsdale, NJ: Erlbaum.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*, 608–621 doi:10.3758/s13428-012-0211-3

Martin, D., & Berry, M. (2007). Mathematical foundations behind latent semantic analysis. In T. Landauer, D. McNamara, D. Simon & W. Kintsch (Eds), *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCowan, I., Moore, D., Dines, J., Gatica-Pérez, D., Flynn, M., Wellner, P., . . . Bourlanr, H. (2004). *On the use of information retrieval measures for speech recognition evaluation, IDIAP-RR73*. Martigny, Switzerland: IDIAP.

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., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. Landauer, D. McNamara, D. Simon & W. Kintsch (Eds), *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex, England: Addison Wesley Longman Ltd.

Millis, K. K., Kim, H. J. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments and Computers*, *36*, 213–221.

Monjurul Islam, M., & Latiful Hoque, A. S. M. (2012). Automated essay scoring using generalized latent semantic analysis. *Journal of Computers*, *7*(3):616–626.

Mulder, G., & Sanders, T. J. M. (2012). Causal coherence relations and levels of discourse representation. *Discourse Processes*, *49*, 501–522.

Nakov, P. (2000, August). *Getting better results with latent semantic indexing*. Paper presented at the Students Presentations at the European Summer School in Logic Language and Information (ESSLLI'00) (pp. 156–166), Birmingham, England.

Nakov, P., Popova, A., & Mateev, P. (2001). *Weight functions impact on LSA performance*. Paper presented at Recent Advances in Natural Language Processing, – RANLP 2001, Tzigov Chark, Bulgaria.

Olde, B. A., Franceschetti, D. R., Karnavat, Graesser, A. C., & The Tutoring Research Group (2002). The right stuff: Do you need to sanitize your corpus when using latent semantic analysis? In W. Gray & C. D. Schunn (Eds), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 708–713). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, *8*, 93–133.

Santrock, J. W. (2012). *Psicología de la Educación [Educational psychology]* (4th ed.). Madrid: McGraw-Hill UNED.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw & N. S. Petersen (Eds), *International encyclopedia of education* (Vol. 4, pp. 20–26). Oxford, UK: Elsevier.

Srihari, S., Collins, J., Shihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffer, J. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence*, *172*, 300–324.

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

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, and Computers*, *35*, , 22–31.

Wolfe, M. B. W., Magliano, J. P., & Larsen, B. (2005). Causal and semantic relatedness in discourse understanding and representation. *Discourse Processes*, *39*, 165–187.

Author Biographies

Guillermo Jorge-Botana is a lecturer at the UNED University of Spain. He has experience in the fields of computational psycholinguistic, natural language technologies, Statistical Language Models, psycholinguistic models, event-related potentials and cognition. He is the cofounder of www.semantialab.es.

José M. Luzón is professor in the education & development department of Universidad Nacional de Educación a Distancia (UNED). He has participated in projects related with new technologies, Multimedia Learning, Statistical Language Models. He is the cofounder of www.semantialab.es.

Isabel Gómez-Veiga has obtained PhD in Educational Psychology, in 2001, from UDC (Universidade da Coruña, Spain). She is professor at the department of developmental and educational psychology at UNED (Universidad Nacional de Educación a Distancia, Spain). Her research interests are focused on language comprehension, reasoning, and the relations between these cognitive variables within educational settings and with instructional purposes.

Jesús I. Martín-Cordero is a professor in the department of educational and developmental psychology at the UNED University of Spain. He teaches courses in education psychology and instruction.