Using latent semantic analysis to grade brief summaries: some proposals

Ricardo Olmos and José A. León*

Departamento de Psicología Básica,
Facultad de Psicología,
Universidad Autónoma de Madrid,
Campus de Cantoblanco, 28049 Madrid, Spain
Fax: +34-91-497-52-15
E-mail: Ricardo.olmos@uam.es
E-mail: joseantonio.leon@uam.es
*Corresponding author

Inmaculada Escudero

Departamento de Psicologia Evolutiva,
Facultad de Psicología,
UNED,
28080 Madrid, Spain
E-mail: iescudero@psi.uned.es

Guillermo Jorge-Botana

Departamento de Psicología Básica,
Facultad de Psicología,
Universidad Autónoma de Madrid,
Campus de Cantoblanco, 28049 Madrid, Spain
E-mail: jorgeybotana@gmail.com

Abstract: In this paper, we present several proposals in order to improve the LSA tools to evaluate brief summaries (less than 50 words) of narrative and expository texts. First, we analyse the quality of six different methods assessing essays that have been widely employed before (Foltz et al., 2000). The second objective is to analyse how new algorithms inspired by some authors (Denhière et al., 2007) that try to emulate human behaviour to improve the reliability of LSA with human graders when assessing short summaries, compared with standard LSA use in expository text. Finally, we present an assessment method to combine LSA as a semantic computational linguistic model with ROUGE-N as a lexical model, to show how combining different automatic evaluation systems (LSA and ROUGE) can improve the quality of assessments in different academic levels.

Keywords: latent semantic analysis; LSA; assessment summaries; ROUGE-N; automated scoring of summaries; brief summary; vector length; expository text; narrative text.

Biographical notes: Ricardo Olmos received his PhD in Behavioural Sciences Methodology and Reading Comprehension (Universidad Autónoma de Madrid). He has research experience in computational linguistics (latent semantic analysis model of computational linguistics). He is an Associate Professor at IE University teaching psychometrics, and last year gave a degree course Introduction to Psychometrics at Universidad Autónoma de Madrid. His fields of expertise are statistics, psychometrics and computational linguistics.

Jose A. León is a Professor of Text Comprehension, Knowledge Acquisition and Reading Comprehension at the Universidad Autónoma de Madrid. His work involves the theoretical study of cognitive processes in comprehension, memory and language, as well as the application of cognitive principles to educational practice. His current research covers a variety of topics including text comprehension, understanding and evaluating inferences, neurocognitive processes, developing automatic scoring technology and computational models using latent semantic analysis.

Inmaculada Escudero is Professor and Head of Research for the Department of Applied Linguistics at Antonio de Nebrija University. She has participated in many public and privately-funded research projects in the fields of technology, reading and comprehension, usability, computational linguistics and sociolinguistics.

Guillermo de Jorge-Botana is a Doctor of Cognitive Science (Universidad Autónoma de Madrid) specialising in latent semantic analysis (LSA) techniques. He works in the technology sector, involved in development and programming as well as analysis and design. He programmes and designs applications in Prosodie for Self-Service and IVR platforms.

1 Introduction

One of the most important aspects of learning and comprehension processes is the assessment of the knowledge acquired by the learner or the comprehender. For decades, educators have often used written material such as essays or summaries to evaluate it. From the constructivist perspective, building explanations or producing written material such as summaries produces and improves comprehension of texts (Graesser et al., 2004). Several studies have demonstrated the importance of knowing how to summarise succinctly in understanding and learning, and have shown it to be a good measure of comprehension processes (Brown et al., 1983; Sherrard, 1985, 1989; Wade-Stein and Kintsch, 2004).

Summaries enjoy a privileged role in research on comprehension of texts and its evaluation. For some authors, a text has not been understood if the reader cannot summarise it (Palinscar and Brown, 1984). While 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 (León et al., 2006). For authors such as van Dijk and

Kintsch (1983), summarising involves the capacity to generalise, synthesise and write coherently. It thus goes far beyond reading, since it implies profound comprehension of what is read. In their model of comprehension, summarising is essential to understanding, since it involves extracting the main content of what is read, at the same time eliminating superficial details. Kintsch (2002) himself used the latent semantic analysis (LSA) model in an attempt to find the phrases that best summarise 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). Summarising, then, involves establishing relationships between important concepts, and presenting them in a coherent, organised 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. Summarising allows us to build on information in the classroom further and better than 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 et al., 2000) that learning to summarise 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. Usually, an evaluation of summarisation involves human judgments of different quality metrics, such as coherence, conciseness, grammaticality, readability, and content (Mani, 2001). However, even simple manual evaluation of summaries on a large scale over a few linguistic quality questions and content coverage as in the Document Understanding Conference (DUC) (Over and Yen, 2003) would require over 3,000 hours of human efforts. This is very expensive and difficult to conduct on a frequent basis. There are tools currently available which can evaluate texts reliably automatically as well as automated essay scoring (Burstein, 2003; Burstein et al., 2003; Dikli, 2006; Elliott, 2003; Higgins et al., 2004; Landauer et al., 1998; Larkey, 1998; Lin, 2004). These tools are apt for awarding a final grade, and more importantly may be valuable for monitoring a student's progress, or to provide students with longitudinal information regarding their level of ability. Two of these automatic tools are ROUGE-N (Lin, 2004) and LSA (Landauer et al., 1998b). Our research has focused mainly on LSA, which we will describe briefly.

LSA offers a mathematical representation of a semantic domain. It can also be conceived as an automatic statistical method for representing the meaning of words and passages of text (Landauer and Dumais, 1997). This tool is capable of analysing a huge dimensional matrix where each row represents a digitalised word (term) and each column represents one paragraph (document). After that, LSA reduces the original matrix via SVD – a mathematical technique that reduces the dimensionality of a matrix into a new semantic space where each word and each document are represented as a single vector. It has been repeatedly demonstrated that this reduced semantic space preserves the semantic relations between words and documents, as humans do. In this semantic space, it is possible to compare units of a piece of information with adjoining units of the text to determine the degree to which the two are semantically related. The units of textual information may be sentences or paragraphs (e.g. Foltz, 1996; Landauer, 1998; Landauer and Dumais, 1997; Landauer et al., 1998a) or summaries (e.g. Foltz, 1996; Kintsch et al., 2000; León et al., 2006). LSA measures the similarity between two pieces of text using the cosine between the two vectors.

During the last 20 years, its capacity to simulate aspects of human semantics has been widely demonstrated (Landauer and Dumais, 1997). 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 et al., 2004; Millis et al., 2007). LSA is an appropriate 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, while 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 to predict the level of comprehension, to evaluate the predominance of reading strategies, or to give appropriate feedback to students, coaching them towards better use of strategies (McNamara et al., 2007; Millis et al., 2007).

Another educational application of LSA uses computer tutors (Graesser et al., 2005; Wade-Stein and 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 Kintsch (2004). This tool coaches during the process of writing a summary. The basic idea underlying this tool is how to teach students to summarise. These computer tutors teach students to resolve problems (e.g., summarising) without individual attention from a teacher, a difficult task if we consider the lack of educational resources available today. In simple terms, what LSA does is compare what students write with texts built into the tools (ideal summaries, main topics, keywords, 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 include 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 et al., 2007).

A third example of an LSA application in the field of education is based on automated assessors (Foltz et al., 1999; Landauer et al., 1998a). 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 uses an expert summary (normally by a grader or teacher) as a source text, thus creating what they call a 'golden summary' (Landauer et al., 1998b; León et al., 2006). These tools automatically provide an essay score, sometimes offering impressive results, proving as reliable as the expert judges (trained graders or teachers) themselves. Another similar application by French authors Dessus and Lemaire (2002) is the APEX system. This system, based on LSA, provides texts for students to summarise, and then evaluates the summary.

2 Objectives

In this paper, we present several proposals for improving LSA tool's evaluation of brief summaries. We focus on a variety of studies where we have tested LSA's assessment,

using short summaries (less than 50 words) of narrative and expository texts. We present these studies in a coherent manner to show the improvement in assessment quality obtained over several years by selecting the best methods in the literature, creating new algorithms that capture some authors' ideas, and finally combining LSA with other automatic system as ROUGE-N (Lin, 2004).

It is important to note that 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 50 words. When texts contain fewer than 1,000 words, it is only natural that the summaries are shorter than those used in most LSA applications (Wade-Stein and Kintsch, 2004; Landauer et al., 1998b, Rehder et al., 1998). The summaries analysed in the three studies presented here all have a maximum length of 50 words. We did not, then, focus on whether LSA is a good system for assessing the quality of essays since there is ample evidence of this. Rather, we have focused on whether LSA can be considered a reliable system for assessment of very short texts.

We present three main objectives in this paper.

1   We analyse the quality of six widely-used methods for assessing essays (Foltz et al., 2000; Landauer and Dumais, 1997), and try to establish which methods should be used when LSA analyses the quality and content of short summaries (León et al., 2006).

2   The second objective is to analyse how new algorithms could improve the reliability of LSA with human graders when assessing short summaries of an expository text (Olmos et al., 2009). In this study, we have been inspired by some authors' ideas that consider LSA in a new light (Denhière et al., 2007; Hu et al., 2007; McNamara et al., 2007; Kintsch, 2001, 2002).

3   Finally, we present an assessment method to combine LSA as a semantic computational linguistic model with ROUGE-N as a lexical model, to show how combining different automatic evaluation systems can improve the quality of assessments at different academic levels.

3   Componential vs. holistic: Which is the best LSA method?

In order to establish a consistently reliable LSA assessment method for grading short texts, we tested LSA's ability to simulate human judgements about summaries. We compared six methods applied by other researchers in previous studies (e.g. Foltz et al., 2000; Kintsch et al., 2000; Landauer and Dumais, 1997). These researchers have distinguished between holistic (H) and componential or analytic methods (C); all of these methods except one have been used previously to score essays. Holistic and componential methods differ in the way they score the summaries. Whereas holistic methods provide a scoring of the summaries on the basis of their overall similarity to the global text (or expert summary), componential methods calculate scores based on the similarity of multiple components of the summary (such as individual sentences, coherence, content or main topics) to the global text. According to Foltz et al. (2000), each approach has its own advantages. Whereas the holistic method can typically provide a more accurate measure of the overall quality of a summary, the componential scoring

method can provide more specific detail about which components of the summary scored better. In this study we selected six different methods, four holistic and two componential, which are described below.

Method H1: Summary-text. This holistic method consists of comparing each student's summary with the full text to derive the LSA cosine. The higher the cosine between the summary and the original text, the better the summary will score. This method has been applied by Kintsch et al. (2000) using summarisation tasks in their summary street computerised tutoring system.

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

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

Method H4: Pregraded summary-ungraded summary. In this final holistic method, a sample of summaries was first graded by 100 instructors, then the cosine between each pregraded summary and the remaining ungraded summaries was computed. Once the cosine was computed, each ungraded summary was assigned the average score of a set of ten very similar summaries, weighted by their similarity. The main strength of this method is that it uses human judgements as the starting point. This method has been applied by Landauer et al. (1998b) to student essays.

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

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

The Spanish LSA corpus used in this study contains 2,059,234 documents (i.e., paragraphs), which include 1,661,954 different terms (without syntax parsing), with the corpus finally set at 337 dimensions. Three hundred and ninety 14 to 16-year-old students from secondary schools in Madrid participated voluntarily in this study. The participants were required to write a concise, four-line summary with a maximum of 50 words (198 summaries of a narrative text and 192 of an expository text). To compare the quality of the LSA assessments, four PhD students evaluated the content of the summary on a scale

of 0 to 4 on the basis of its four main components, and the summary coherence on a scale of 0 (incoherent) to 6 (highly coherent).

Interrater reliability between human expert ratings ranged from .79 to .86 (Pearson correlation). The reliability between LSA and humans raters was statistically significant for all six methods, but in general the holistic methods work better than the componential methods (especially for the expository summaries).

The results showed that the two best methods were pregraded summary-ungraded summary, and the summary-expert summaries. One important finding of our study concerns whether the length of a summary (maximum 50 words) is a key factor in assessing quality in relation to its LSA cosine. For this aspect, our correlations were similar to those found by Kintsch et al. (2000) in their study of narrative texts on ancient civilisations. One interpretation of our results would be that restrictions on text length are compensated for by a greater conceptualisation of the summary and by a higher concentration of key information or main topics contained in the texts. This viewpoint supports the idea that LSA is sensitive to semantic information in terms of conceptualisation and abstraction. The reliability of these two methods compared with human graders ranges from .41 to .63 (Pearson correlation) for the expository text and .46 to .58 for the narrative text. In these two methods evaluations were made using only information contained in the summaries. The three worst methods directly compared information based on the text, instead of an ideal or 'golden' text. However, the philosophy of the summary-expert summaries method is better for the main aim of LSA (automatic assessment) because it does not need any prior information from human graders (unlike the pregraded-ungraded method, which also works particularly well, but for which LSA requires a pool of summaries previously scored by human graders). Also, the summary-expert summaries method underlines the key idea that it is a good idea to create some ideal texts as reference texts to compare with the candidate texts (in our case the student summaries).

4   New algorithms for evaluating short summaries of expository texts

In the previous study, we analysed the LSA evaluation methods that have been used in recent years (Foltz et al., 2000; Kintsch et al., 2000; Landauer and Dumais, 1997). One of the limitations of these methods has been to conceive LSA as a model of semantic *representation*. In this sense, LSA is a static theory of language where each word is represented as a vector in a reduced and latent semantic space. However, some authors have tried to give LSA more psychological plausibility, two excellent examples being studies by Denhière et al. (2007) and Kintsch (2001). These authors have wondered what would happen if we were to conceive LSA as a model of semantic *processing*. In the study by Denhière and his colleagues, LSA is conceived as a semantic space that models children's semantic memory. Kintsch's study describes the predication algorithm, which makes language more context-dependent, and among other things this reduces the impact of the polysemy problem inherent in LSA (Deerwester et al., 1990). Another limitation of previous studies is the mathematical information used in the methods analysed. Some authors have therefore proposed new mathematical extensions (Hu et al., 2007; McNamara et al., 2007). The study by Hu et al. presents the 'adaptive method' to make more efficient use of the latent semantic space when LSA is assessing physics protocols by students. Another example is presented in a study by McNamara et al. using a

weighting algorithm to reflect variations in the importance words have in different sentences. With this contribution in mind, the objective was to analyse how new algorithms could improve the reliability of LSA compared to human graders assessing short summaries of an expository text. We propose the introduction of new algorithms in this study to reflect the mathematical-psychological models proposed by the above authors in the context of summary assessment. The algorithms are as follows.

The common semantic network algorithm. LSA is not capable of distinguishing multiple senses of a word, since a single vector represents only one word. This is known as LSA's *polysemy problem* (Deerwester et al., 1990). Kintsch (2001) showed that the LSA model could be used to provide a good semantic representation (the algorithm is applied to sentences with the structure *argument-predicate*), as long as the specific role of the predicate is taken into account. The essence of the algorithm is to strengthen aspects of the predicate appropriate to the argument. In other words, this algorithm extracts a context-dependent meaning; for example, in LSA's representation of the sentence 'this lawyer is a shark', Kintsch (2000, 2001) proposed that only neighbours of the predicate associated with the context need be considered. Therefore, associated neighbours such as *aggressive*, *predatory* or *tenacious* would be activated, but not *fish*, *swimmer* or *gills*, since although they are close neighbours of literal shark-properties, they are not related to the argument. In this way, the algorithm incorporates information about neighbours, so that the information in *shark* is linked via a semantic network to *lawyer*. We used the same idea to establish a common semantic network between the summary by a student and the summarised text. The general idea behind our adaptation of this algorithm was to provide additional semantic information in the summary vector. Thus for a summary, instead of representing the vector with the sum of its words we added to the summary its closest neighbours, expanding the semantic network. Now, the summary comprised its own words and others semantically related to it. In psychological terms, the algorithm means that when we express something in a piece of language, the meaning conveyed is more than that expressed explicitly. Therefore, our final vector represented in the LSA space consists of the words from the summary and the concepts that are semantically most closely related. We adapted this algorithm because instead of adding to the student summary its *n* closest neighbours, we only added to each student's summary the *p* terms most closely related to the expository text (where *p* < *n*). The common semantic network algorithm first extracted the summary's 50 closest neighbours. In the semantic network, we then included not the 50 (*n*) neighbours but the 20 (*p*) most closely related to the expository text, where *p* < *n* following Kintsch's (2007) criteria. Thus, 20 terms were added to the vector summary, and these terms were semantically related to both the summary and the expository text. The semantic network was extended by

1   activating the summary's closest neighbours

2   suppressing neighbours not related to the text

3   retaining those neighbours with relatively strong links to the text.

The Best Dimension algorithm. Instead of using the entire semantic space to represent a text, Hu et al. (2007) used only dimensions that best contribute to improving LSA assessment's verbal protocols. This algorithm makes intelligent, discriminative use of the semantic space. In our case, we applied it to summaries by randomly selecting 30 out of 192 summaries to train the method, and the remaining 162 to validate it, in order to avoid

overfitting the algorithm to the whole sample. The four graders rated them and we noted the average grade for each summary; we thus previously obtained 30 ratings by human graders. LSA rated these 30 summaries as follows. First, it removed the dimension that obtained the worst Pearson correlation between LSA and human grader average, and obtained the best p–1 dimension semantic space in terms of LSA-human grader reliability. Second, it removed the dimension in this reduced space that most reduced LSA-human grader reliability. The algorithm continued until the worst 20% of the dimensions were removed, leaving us with the semantic space that most contributed to LSA-human grader reliability. Each summary vector therefore had 80% of the original information and LSA used only the most relevant features of the semantic space, in much the same way as human graders consider only the most relevant features when assessing summaries.

The Euclidean distance. The cosine has habitually been the measure used by LSA to evaluate texts (or rather to evaluate similarity between texts). As a consequence, we noticed that some summaries with insufficient detail were rated as very similar to expert summaries or to the summarised text, i.e., ratings provided by LSA were overestimated. For example, suppose that we use the summary-text method (see Section 3, 'The six methods') to assess summaries where we compare the student summary with the text summarised. Figure 1 represents this case graphically – the large arrow represents the vector of the text, and the small arrows represent two student summaries, called S1 and S2. Note that S1 is closer to the text vector than S2 in terms of the cosine (smaller angle), but if we consider the Euclidean distance (the distance between the arrow tips), S2 is closer to the text than S1. As a solution to the cosine problem, we used the Euclidean distance measure (see a description of this and other measures in Rehder et al., 1998). This measure incorporates both vector length and cosine. Vector length reflects the quantity of detail in the summary, and the cosine reflects the amount of semantic similarity. Euclidean distance is therefore an algorithm that contains more information about the summary content, and probably improves the reliability of LSA for assessing summaries. We feel this measure is particularly sensitive to the method used for assessing summaries. Euclidean distance would offer different quality using the summary-text method compared to the summary-expert summaries method. Since an expert summary has approximately the same level of detail as a student summary (same vector length), Euclidean distance would not be an appropriate or sensitive measure with the latter method.

The Spanish LSA database developed for this study contains 372 documents with similar contents to the expository text used in the study. These documents were taken from internet resources, textbooks, and online encyclopaedias. In total, it holds 5,995 lemmatised words. The semantic space was set at 75 dimensions, which cover 40% of the total variance. The expository summaries were the same as those used in the previous study, and these 192 summaries were rated by four expert graders on a 0–10 point scale. As in the previous study, the expert graders' ratings are used to establish LSA reliability with each of the stated algorithms (common semantic network, Best Dimension and Euclidean distance) and with standard LSA procedure. The rating using standard LSA procedure was taken as the baseline used for comparing the other algorithms. Each algorithm derives its own vectors for each summary and the standard has the usual vector for each summary. To rate each summary, we compared each vector with the text vector (summary-text method) or with six expert summaries (the summary-expert summaries method).

Figure 1   An example of a cosine measure overestimating a summary score
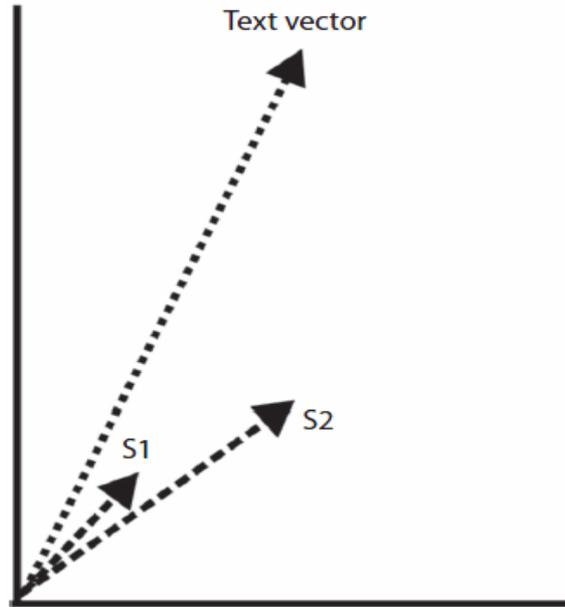
Results. We found differences in the magnitude of reliabilities, depending on the algorithm [F (3,24) = 11.50, MSE = 0.02, p = .05]. However, these results are modulated by the interaction effect (see Figure 2).

Figure 2   Interaction effect between algorithm and method
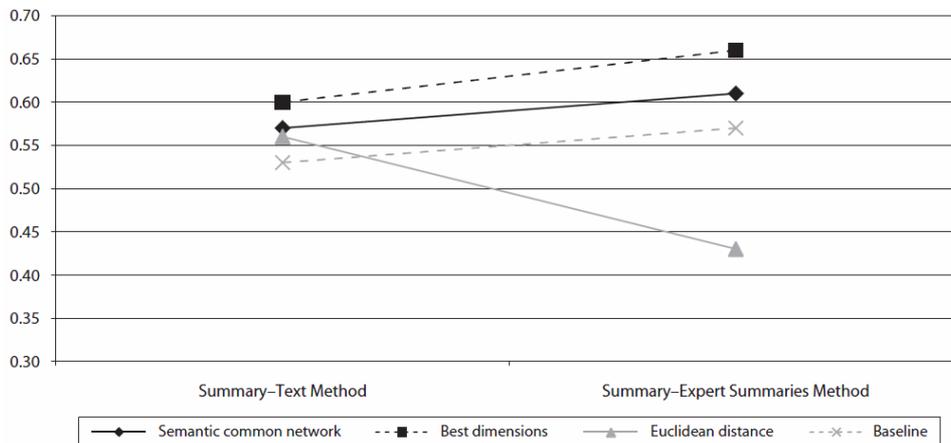
There was an interaction effect between algorithm and method [F (3,24) = 7.21, MSE = 0.02, p = .05]. The interaction was caused by the Euclidean algorithm: failing to increase the mean reliability in the expert summaries method, this algorithm does not seem to work well. The simple interaction effects were as follows: we found that the Best Dimension algorithm showed higher reliability than the baseline in the summary-text method and in the summary-expert summaries method (p = .05). We did not find any

other differences in the mean between algorithms for the summary-text method ($p = .05$). The common semantic network algorithm and the baseline algorithm did not differ significantly, although the common semantic network algorithm has higher reliability means with both methods. This statistical test is not very powerful (four cases per group), which probably explains the lack of significant differences. Finally, the reliability of Euclidean distance proved the same as the other algorithms in the summary-text method ($p = .05$), but gave its worst results in the summary-expert summaries method ($p = .05$).

Our objective was to find algorithms closer to the dynamic cognitive processing of texts than standard LSA, which gives a static representation of semantics. In order to improve LSA performance, some authors have proposed extensions. Some of these extensions have concentrated on algorithms that select dimensional information intelligently (Hu et al., 2007), or on algorithms that change the static semantic representation into a context-dependent representation (Kintsch, 2008). Our solution has been to create three new algorithms that incorporate (or remove) some adaptive information in the vector representation of the text. Thus, the Best Dimension algorithm suppresses those dimensions that affect reliability. The common semantic network adds semantically related neighbours connected with the summarised text to the vector summary; Euclidean distance incorporates vector length as a measure. Are these algorithms capable of improving assessment quality for expository texts? Of the three algorithms used in this study, only the Best Dimension algorithm supports this idea. In LSA, dimensions have no explicit interpretation (which is not the case for factorial analysis), but not all the semantic dimensions are task-relevant. In the same way, we probably do not use all our semantic memory when we undertake a task. This algorithm seems to remove some dimensions that contain noise in assessing written material, and retain those dimensions that discriminate clearly between good and bad summaries. In the near future, the next step could be to rotate the semantic space to find a new base with meaningful dimensions (Hu et al., 2007). The common semantic network algorithm also showed promising trends and results; it is based on the idea of simulating a number of semantic phenomena, one of which is context dependency in similarity assessment (others mentioned by Kintsch, 2007, are metaphor comprehension or causal inferences). This algorithm enriches vector summaries with relevant terms, but in the future it should be refined – for example by enriching the summary only if its neighbours exceed a fixed threshold. Thus, an anomalous summary might not benefit from the algorithm, while summaries scoring high semantically would. Euclidean distance did not obtain as high LSA-human grader reliability in the summary-expert summaries method as in other studies (Jorge-Botana et al., 2010), but since it incorporates vector length it would be a good measure for certain tasks (e.g., we have recently seen that the Euclidean distance algorithm can distinguish better between expert and novice answers than the cosine can). We think this method was inappropriate for the summary-expert summaries method because the Euclidean distance cannot discriminate well between good and bad summaries. In this method, LSA compares the Euclidean distance between an expert summary and a student summary. Probably the measure of the distance between expert and student summaries is not sufficiently sensitive and cannot provide good ratings, since both groups are forced to keep within a maximum length of 50 words – we believe this to be the most plausible explanation for the interaction found with this algorithm. The method works well when at least some of the texts compared are not limited in length, as we can see in the summary-text method.

5 Combining LSA with ROUGE-N to grade brief summaries

The last study focuses on another strategy that is probably most fruitful in automatic evaluation of essays. The strategy is simple: combining different automatic evaluation systems to improve the assessment quality of essays. In our study, we use a combined method that captures semantic similarity and semantic space (LSA) as well as lexical similarity (ROUGE-N), analysing its reliability comparing to human graders' evaluations. The information that we used from LSA was semantic similarity commonly provided by the cosine measure, together with information on the extent of knowledge LSA has of the terms represented in a semantic space (provided by the vector length measure).

The novelty in this study is the incorporation of the lexical measure ROUGE-N (Lin, 2004). The typical approach to finding the similarity between two text segments is to use a simple lexical matching method, and produce a similarity score based on the number of lexical units that occur in both input segments. ROUGE-N includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between a candidate summary to be evaluated and the ideal summaries (Lin, 2004). For this study, we selected ROUGE-N because it included an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

$$ROUGE\ N = \frac{\sum_{S \in \{referencesummary\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{referencesummary\}} \sum_{gram_n \in S} Count(gram_n)}$$

where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. It is clear that ROUGE-N is a recall-related measure because the denominator of the equation is the total sum of the number of n-grams occurring on the reference summary side. Note that the number of n-grams in the denominator of the ROUGE-N formula increases as we add more references. This is intuitive and reasonable because there may be many good summaries.

The main aim of this study was to use an LSA-based computational method and ROUGE-N to reliably evaluate particularly brief summaries. The LSA method incorporates the essential information from the latent semantic space:

1 a measure of semantic similarity using the cosine

2 a measure of the vector length or extent of knowledge about the text

3 ROUGE-N incorporates the lexical information with the co-occurrences between student summaries and reference summaries.

786 Spanish students from four grade levels took part in this study. Student ages ranged from 10 to 23 years. Of the 786 students, 238 were from 6th grade, followed by 192 students from 8th grade, 198 students from 10th grade, and lastly 158 university students. Of the 786 summaries they produced, 396 were summaries of a narrative text, and the remaining 390 were summaries of an expository text. Again, we compare the quality of

our automatic method with the evaluation of four trained PhD students that evaluated the 786 summaries on a scale of 0 to 10. LSA was trained with a general domain Spanish corpus that had 2,059,234 documents (i.e., paragraphs) and 1,661,954 different terms. A semantic space with 337 dimensions was used.

5.1   The proposed method

To implement our method, we used a database comprising 100 summaries of narrative text and 100 expositive text 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 and ROUGE-N, although it is not used to evaluate the reliability of the method. 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 an ordinary least squares linear regression 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:

$$NarrativeScore = \beta_0 + \beta_1 * Vectorlength + \beta_2 * Similarity + \beta_3 * ROUGE \cdot N$$

And the regression equation for the expository text was:

$$ExpositoryScore = \beta_0 + \beta_1 * Vectorlength + \beta_2 * Similarity + \beta_3 * ROUGE \cdot N$$

where β_0 is the constant, β_1 is the coefficient for vector length, β_2 is the coefficient for semantic similarity, and β_3 is the coefficient for the ROUGE n-gram measure. Once the regression calculations are done, we took a new sample of summaries (the validation sample). This time there were 296 summaries of the narrative text and 290 summaries of the expositive 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, which was then used to assess the reliability of the scores awarded by LSA and ROUGE-N using the regression equations. The independent validation sample was used to avoid overfitting, to make it easier to generalise to new summaries.

Results: The regression line to predict the grades for the summaries of the narrative text was:

$$NarrativeScore = -.73 + 4.18 * Vectorlength + 6.69 * ExpertMethod + .69 * ROUGE$$

where the vector length coefficient was statistically significant (T = 2.73, p < .05; 95% CI: 1.14-7.22), as too was the expert method coefficient of semantic similarity (T = 3.08, p < .05; 95% CI: 2.38-11.00), and the ROUGE-N coefficient (T=2.96, p< .05; 95% CI: .23-1.15).

The regression line obtained for the expository text was:

$$ExpositoryScore = -4.31 + 10.36 * Vectorlenght + 11.44 * ExpertMethod + \\ .73 * ROUGE$$

Once again, the vector length coefficient was statistically significant (T = 7.61, p < .05; 95% CI: 7.66-13.06), as was the coefficient associated with expert method (T = 6.33, p < .05; 95% CI: 7.85-15.03), and the ROUGE-N coefficient (T=4.21, p<.05; 95% CI: .38 – 1.07).

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

The LSA and ROUGE-N-grader reliabilities were calculated using the validation sample. This sample was set aside to avoid overfitting of reliabilities, and thus allow generalisation of results to other summaries. Table 1 shows the reliability of LSA and ROUGE-N with individual judges and with the average judges' scores in both texts (calculated with Pearson's r correlation). For the narrative text, the reliability of LSA and ROUGE-N ranged from .65 to .70 for the individual judges, and reached .72 with the average judges' scores. As for the expository text, the LSA and ROUGE-N-grader reliability ranged from .76 to .83, reaching .85 with the average judges' scores. The results for the reliability scores were better for expository than narrative text.

Table 1 LSA+ROUGE-N-grader reliability for each text and human grader

Text	Human grader				
	Grader 1	*Grader 2*	*Grader 3*	*Grader 4*	*Grader average*
Narrative text	.65**	.70**	.66**	.67**	.72**
Expository text	.76**	.80**	.82**	.83**	.85**

Note: **p < .01

We found significant differences between LSA and ROUGE-N-human grader reliability for the expository text (r = .85, p < .01) and LSA and ROUGE-N-human grader reliability for the narrative text (r = .72, p <. 01). Reliability for the expository text was thus significantly higher (p <. 01).

Table 2 Unique contributions and combined effects of the three effects considered (LSA cosine and vector length) and ROUGE-N on grader assessments

Text		Unique effects			Combined effects
		LSA (semantic similarity)	*LSA (vector length)*	*ROUGE-N*	*LSA (semantic similarity and vector length) + ROUGE-N*
Narrative	R square	.37	.29	.46	.52
	Reliability (Pearson correlation)	.61**	.54**	.68**	.72**
Expository	R square	.46	.49	.52	.72
	Reliability (Pearson correlation)	.68**	.70**	.72**	.85**

Note: **p<.01

To study the effect of combining LSA and ROUGE-N it is important to see the individual effects on predicting grader assessments. Thus, we ran three regression equations isolating each of the effects considered in this study. First, we use the LSA semantic similarity (expert method) to explain the graders assessment. Secondly, we included only vector length measure in the regression equation of the LSA measure. Finally, we run the regression with the ROUGE-N measure. Table 2 shows these unique contributions to explaining grader assessments.

For the narrative text, the most important effect is the ROUGE-N measure (R square = .46). However, there is an advantage when the three effects are included simultaneously, as we can see in the combined effects column. When the three effects are in the equation, the change in the proportion of variance is statistically significant, with an improvement of about 4%. In fact, in a *stepwise* regression, the first effect that enters in the equation to predict the grader assessments is ROUGE-N. In a second statistically significant model, the next effect incorporated in the equation is the LSA semantic measure (expert method). Finally, a third model is run in which vector length improves the model significantly. For the expository text the individual effects have approximately the same importance in predicting the grader assessments (R square varied between .46 and .52). The combined effects column shows a considerable benefit when the three effects are included simultaneously (the gain in the proportion of variance is more than 20%). Thus, the combination of the three effects seems more important for the expository text than in the narrative text

LSA significantly increased the reliability when combined with ROUGE-N. This integrated method improves the model significantly, and the reliability seems more important for the expository text than the narrative text (Pearson r correlation .72 for narrative, and .85 for expository text). A possible interpretation of these results supports the idea that a method combining semantic and lexical similarities increased reliability compared with human graders. The results showed progress toward fulfilling these aims, although in general they are relatively more satisfactory for evaluations of the expository summaries studied in this paper. We must assume, however, that our results do not constitute a sufficient basis for the generalisation of our findings, since they were drawn from only two text samples (one narrative and one expository). Further research is needed.

6   General conclusions

The availability of automatic tools to evaluate reliably and help detect strong or weak points in 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 in recent years, and one of the fastest-growing areas of application has been the field of education. This tool needs to be complemented with new algorithms to overcome some of its limitations (Jorge-Botana et al., 2009; Olmos et al., 2009), and mathematically optimise the usage of the latent semantic space (Hu et al., 2007). It is also possible to combine several of these algorithms, as in our implementation combining cosine and vector length. We should, however, link them with psychological models such as semantic memory (Denhière et al., 2007; Jorge-Botana et al., 2010) or with other computational models of language (Steyvers and Griffiths, 2007). Once all of these

contributions are added, the potential and the capability of LSA in the educational sector will be far greater.

The studies also present 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, 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 generalisation of our findings. In general, our studies showed a strong dependency related to the number of words contained in the summary. For very short summaries, the quality of evaluations made using these methods as well as by human graders depends entirely on the number of words in the summary – a serious limitation to evaluating very brief texts that can be partially overcome using the methods we propose. More research is required, however, in order to improve assessment quality in these highly conceptualised texts.

Acknowledgements

References

Bransford, J.D., Brown, A.L. and Cocking, R.R. (2000) *How People Learn: Brain, Mind, Experience, and School*, National Research Council Commission on Behavioral and Social Sciences and Education, National Academy Press, Washington, DC.

Brown, A.L., Bransford, J.D., Ferrara, R.A. and Campione, J.C. (1983) 'Learning, remembering, and understanding', in Flavell, J. and Markman, E.M. (Eds.): *Handbook of Child Psychology Cognitive Development*, 4th edition, Vol. 3, pp.515–629, Wiley, New York.

Burstein, J. (2003) T'he e-rater® scoring engine: Automated essay scoring with natural language processing', in anonymous (Eds.): *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Burstein, J., Chodorow, M. and Leacock, C. (2003) 'CriterionSM: online essay evaluation: an application for automated evaluation of student essays', in *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*', Vol. 41, pp.391–407.

Denhière, G., Lemaire, B., Bellissens, C. and Jhean-Larose, S. (2007) 'A semantic space modeling children's semantic memory', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.143–167, Erlbaum, Mahwah, NJ.

Dessus, P. and Lemaire, B. (2002) 'Using production to assess learning: an ILE that fosters self-regulated learning', *International Conference on Intelligent Tutoring Systems (ITS'2002)*, Springer Verlag, Berlin, LNCS 2363, pp.772–781.

Dikli, S. (2006) 'An overview of automated scoring of essays', *Journal of Technology, Learning, and Assessment*, Vol. 5, No. 1, pp.3–35, available at http://www.jtla.org.

Elliott, S. (2003) 'Intellimetric: from here to validity', in Shermis, M. and Burstein, J. (Eds.): *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Foltz, P. W. (1996) 'Latent semantic analysis for text-based research', *Behavior Research Methods, Instruments and Computers*, Vol. 28, No. 2, pp.197–202.

Foltz, P.W., Gilliam, S. and Kendall, S. (2000) 'Supporting content-based feedback in on-line writing evaluation with LSA', *Interactive Learning Environments*, Vol. 8, pp.111–128.

Foltz, P.W., Laham, D. and Landauer, T.K. (1999) 'The intelligent essay assessor: applications to educational technology', *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, Vol. 1, available at http://knowledge-technologies.com (accessed on 29 June 2004).

Graesser, A.C., Chipman, P., Haynes, B.C. and Olney, A. (2005) 'AutoTutor: an intelligent tutoring system with mixed-initiative dialogue', *IEEE Transactions in Education*, Vol. 48, pp.612–618.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H, Ventura, M., Olney, A. and Lowerse, M.M. (2004) 'AutoTutor: a tutor with dialogue in natural language', *Behaviour Research Methods, Instruments, and Computers*, Vol. 36, pp.180–193.

Graesser, A.C., Penumatsa, P., Ventura, M., Cai, Z. and Hu, X. (2007) 'Using LSA in AutoTutor: learning through mixed-initiative dialogue in natural language', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.243–262, Erlbaum, Mahwah, NJ.

Higgins, D., Burstein, J., Marcu, D. and Gentile, C. (2004) 'Evaluating multiple aspects of coherence in student essays', in *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Boston, MA.

Hu, X., Cai, Z., Wiemer-Hastings, Graesser, A.C. and McNamara, D.S. (2007) 'Strengths, limitations, and extensions of LSA', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.401–426, Erlbaum, Mahwah, NJ.

Jorge-Botana, G., León, J.A., Olmos, R. and Escudero, I. (2010) 'Latent semantic analysis parameters for essay evaluation using small-scale corpora', *Journal of Quantitative Linguistics*, Vol. 17, No.1, pp.1–29.

Jorge-Botana, G., Olmos, R. and León, J.A. (2009) 'Using LSA and the predication algorithm to improve extraction of meanings from a diagnostic corpus', *Spanish Journal of Psychology*. Vol. 12, No. 2, pp.424–440.

Kintsch, E., Steinhart, D., Stahl, G. and LSA Research Group (2000) 'Developing summarization skills through the use of LSA-based feedback', *Interactive Learning Environments*, Vol. 8, No. 2, pp.87–109.

Kintsch, W. (2000) 'Metaphor comprehension: a computational theory', *Psychonomic Bulletin & Review*, Vol. 7, No. 2, pp.257–266.

Kintsch, W. (2001) 'Predication' *Cognitive Science*, Vol. 25, pp.173–202.

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

Kintsch, W. (2007) 'Meaning in context', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.89–106, Erlbaum, Mahwah, NJ.

Kintsch, W. (2008) 'Symbol systems and perceptual representations', in de Vega, M., Glenberg, A. and Graesser, A. (Eds.): *Symbols, Embodiment and Meaning*, pp.145–164, University Press, Oxford.

Landauer, T.K. (1998) 'Learning and representing verbal meaning: the latent semantic analysis theory', *Current Directions in Psychological Science*, Vol. 7, pp.161–164.

Landauer, T.K. and 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*, Vol. 104, pp.211–240.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998a) 'Introduction to latent semantic analysis', *Discourse Processes*, Vol. 25, pp.259–284.

Landauer, T.K., Laham, D. and Foltz, P.W. (1998b) 'Computer-based grading of the conceptual content of essays', unpublished manuscript.

Larkey, L. (1998) 'automatic essay grading using text categorization techniques' in *Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp.90–95.

León, J.A., Olmos, R., Escudero, I., Cañas, J.J. and 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*, Vol. 38, No. 4, pp.616–627.

Lin, C.Y. (2004) 'Looking for a few good metrics: ROUGE and its evaluation', in *Proceedings of NTCIR Workshop 2004*, Tokyo, Japan.

Mani, I. (2001) *Automatic Summarization*, John Benjamins Publishing Co.

McNamara, D., Boonthum, C., Levinstein, I. and Millis, K.K. (2007) 'Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.227–241, Erlbaum, Mahwah, NJ.

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

Millis, K.K., Magiano, J.P., Wiemer-Hastings, K., Todaro, S. and McNamara, D. (2007) 'Assesing and improving comprehension with latent semantic analysis', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.207–225, Erlbaum, Mahwah, NJ.

Olmos, R., León, J.A., Jorge-Botana, G. and Escudero, I. (2009) 'New algorithms assessing short summaries in expository texts using latent semantic analysis', *Behaviour Research Methods, Instruments, and Computers*, Vol. 41, pp.944–950.

Over, P. and Yen. J. (2003) 'An introduction to DUC 2003 – intrinsic evaluation of generic news text summarization systems'.

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

Rehder, B., Schreiner, M.E., Wolfe, B.W., Laham, D., Landauer, T.K. and Kintsch, W. (1998) 'Using latent semantic analysis to assess knowledge: some technical considerations', *Discourse Processes*, Vol. 25, pp.337–354.

Sherrard, C. (1985) 'The psychology of summary writing: applying text linguistics', *Journal of Technical Writing and Communication*, Vol. 15, No. 3, pp.247–258.

Sherrard, C. (1989) 'Teaching students to summarize: applying text linguistics', *System*, Vol. 17, No. 1, pp.1–11.

Steyvers, M. and Griffiths, T. (2007) 'Probabilistic topic models', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *The Handbook of Latent Semantic Analysis*, pp.427–448, Erlbaum, Mahwah, NJ.

Van Dijk, T.A. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*, Academic Press, New York.

Wade-Stein, D. and Kintsch, E. (2004) 'Summary street: interactive computer support for writing.', *Cognition and Instruction*, Vol. 22, No. 3, pp.333–362.