# E-Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study

Pete Thomas, Debra Haley, Anne deRoeck, Marian Petre
Computing Research Centre, Department of Computing
The Open University, Walton Hall, Milton Keynes, UK MK7 6AA
P.G.Thomas;D.T.Haley;A.Deroeck;M.Petre [at] open.ac.uk

## Abstract

Latent Semantic Analysis (LSA) is a statistical Natural Language Processing (NLP) technique for inferring meaning from a text. Existing LSA-based applications focus on formative assessment in general domains. The suitability of LSA for summative assessment in the domain of computer science is not well known. The results from the pilot study reported in this paper encourage us to pursue further research in the use of LSA in the narrow, technical domain of computer science.

This paper explains the theory behind LSA, describes some existing LSA applications, and presents some results using LSA for automatic marking of short essays for a graduate class in architectures of computing systems.

## 1 Introduction

This paper describes a pilot study undertaken to investigate the feasibility of using Latent Semantic Analysis (LSA) for automatic marking of short essays in the domain of computer science. These short essays are free-form answers to exam questions - not multiple choice questions (MCQ). Exams in the form of MCQs, although easy to mark, do not provide the opportunity for deeper assessment made possible with essays.

This study employs LSA in several areas that are under-researched. First, it uses very small corpora – less than 2,000 words compared to about 11 million words in one of the existing, successful applications (Wade-Stein & Kintsch, 2003). Second, it involves the specific, technical domain of computer science. LSA research usually involves more heterogeneous text with a broad vocabulary. Finally, it focuses on summative assessment where the accuracy of results is paramount. Most LSA research has involved formative assessment for which more general evaluations are sufficient.

The study investigates one of the shortcomings of LSA mentioned by Manning and Schütze (1999, p. 564). They report that LSA has high recall but low precision. The precision declines because of spurious co-occurrences. They claim that LSA does better on heterogeneous text with a broad vocabulary. Computer science is a technical domain with a more homogeneous vocabulary, which results, possibly, in fewer spurious co-occurrences. A major question of this research is how LSA will behave when the technique is stretched by applying it to a narrow domain.

Section 2 gives the history of LSA and explains how it works. Section 3 describes several existing LSA applications related to e-assessment. Section 4 provides the motivation for more LSA research and reports on a pilot study undertaken to assess the feasibility of using LSA for automatic marking of short essays in the domain of computer science. Section 5 lists several open issues and areas for improvement that future studies will address. Finally, Section 6 summarises the paper.

## 2 What is Latent Semantic Analysis?

“Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text” (Landauer, Foltz & Laham, 1998). It is a statistical-based natural language processing (NLP) method for inferring meaning from a text[1]. It was developed by researchers at Bellcore as an information retrieval technique (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) in the late 1980s. The earliest application of LSA was Latent Semantic Indexing (LSI) (Furnas, et al., 1988; Deerwester, et al., 1990). LSI provided an advantage over keyword-based methods in that it could induce associative meanings of the query (Foltz, 1996) rather than relying on exact matches.

Landauer and Dumais (1997) promoted LSA as a model for the human acquisition of knowledge. They developed their theory after creating an information retrieval tool and observing unexpected results from its use. They claimed that

[1] The researchers originally used the term LSI (Latent Semantic Indexing) to refer to the method. The information retrieval community continues to use the term LSI.

LSA solves Plato's problem, that is, how do people learn so much when presented with so little? Their answer is the inductive process: LSA "induces global knowledge indirectly from local co-occurrence data in a large body of representative text" (Landauer & Dumais, 1997).

From the original application for retrieving information, the applications of LSA have evolved to systems that more fully exploit its ability to extract and represent meaning. Recent applications based on LSA compare a sample text with a pre-existing, very large corpus to judge the meaning of the sample.

To use LSA, researchers amass a suitable corpus of text. They create a term-by-document matrix where the columns are documents and the rows are terms (Deerwester, et al., 1990). A term is a subdivision of a document; it can be a word, phrase, or some other unit. A document can be a sentence, a paragraph, a textbook, or some other unit. In other words, documents contain terms. The elements of the matrix are weighted word counts of how many times each term appears in each document. More formally, each element, $a_{ij}$ in an i x j matrix is the weighted count of term i in document j.

LSA decomposes the matrix into three matrices using Singular Value Decomposition (SVD), a well-known technique (Miller, 2003) that is the general case of factor analysis. Deerwester et. al., (1990) describe the process as follows.

Let t = the number of terms, or rows
d = the number of documents, or columns
X = a t by d matrix

Then, after applying SVD, X = TSD, where

m = the number of dimensions, m <= min(t,d)
T = a t by m matrix
S = an m by m diagonal matrix, i.e., only diagonal entries have non-zero values
D = an m by d matrix

LSA reduces S, the diagonal matrix created by SVD, to an appropriate number of dimensions k, where k << m, resulting in S'. The product of TS'D is the least-squares best fit to X, the original matrix (Deerwester, et al., 1990).

The literature often describes LSA as analyzing co-occurring terms. Landauer and Dumais (1997) argue it does more and explain that the new matrix captures the "latent transitivity relations" among the terms. Terms not appearing in an original document are represented in the new matrix as if they actually were in the original document (Landauer & Dumais, 1997). LSA's ability to induce transitive meanings is considered especially important given that Furnas et. al. (1982) report fewer than 20% of paired individuals will use the same term to refer to the same common concept.

LSA exploits what can be named the transitive property of semantic relationships: If A→B and B→C, then A→C (where → stands for is semantically related to). However, the similarity to the transitive property of equality is not perfect. Two words widely separated in the transitivity chain can have a weaker relationship than closer words. For example, LSA might find that copy → duplicate → double → twin → sibling. Copy and duplicate are much closer semantically than copy and sibling.

Finding the correct number of dimensions for the new matrix created by SVD is critical; if it is too small, the structure of the data is not captured. Conversely, if it is too large, sampling error and unimportant details remain, e.g., grammatical variants (Deerwester, et al., 1990; Miller, 2003; Wade-Stein & Kintsch, 2003). Empirical work involving very large corpora shows the correct number of dimensions to be about 300 (Landauer & Dumais, 1997; Wade-Stein & Kintsch, 2003).

Creating the matrices using SVD and reducing the number of dimensions, often referred to as training the system, requires a lot of computing power; it can take hours or days to complete the processing (Miller, 2003). Fortunately, once the training is complete, it takes just seconds for LSA to evaluate a text sample (Miller, 2003).

## 3 Using LSA for assessment

### 3.1 Types of assessment

Electronic feedback, or e-assessment, is an important component of e-learning. LSA, with its ability to provide immediate, accurate, personalised, and content-based feedback, can be an important component of an e-learning environment.

Formative assessment provides direction, focus, and guidance concurrent with the learner engaging in some learning process. E-assessment can provide ample help to a learner without requiring added work by a human tutor. A learner can benefit from private, immediate, and convenient feedback.

Summative assessment, on the other hand, happens at the conclusion of a learning episode or activity. It evaluates a learner's achievement and communicates that achievement to interested parties. Summative assessment using LSA shares the virtues of formative assessment and can produce more objective grading results than those that can occur when many markers are assessing hundreds of student essays.

The applications described in the next section use LSA to provide formative assessment. Section 4 discusses a pilot study that focuses on summative assessment.

### 3.2 Existing applications

Much work is being done in the area of using LSA to mark essays automatically and to provide content-based feedback. One of the great advantages of automatic assessment of essays is its ability to provide helpful, immediate feedback to the learner without burdening the teacher. This application is particularly suited to distance education, where opportunities for one-on-one tutoring are infrequent or non-existent (Steinhart, 2001). Existing systems include Apex (Lemaire & Dessus, 2001), Autotutor (Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999), Intelligent Essay Assessor (Foltz, Laham & Landauer, 1999), Select-a-Kibitzer (Miller, 2003), and Summary Street (Steinhart, 2001; Wade-Stein & Kintsch, 2003). They differ in details of audience addressed, subject domain, and advanced training required by the system (Miller, 2003). They are similar in that they are LSA-based, web-based, and provide scaffolding, feedback, and unlimited practice opportunities without increasing a teacher's workload (Steinhart, 2001). All of them claim that LSA correlates as well to human markers as human markers correlate to one another. See (Miller, 2003) for an excellent analysis of these systems.

## 4 E-Assessment pilot study

Although research using Latent Semantic Analysis (LSA) to assess essays automatically has shown promising results (Chung & O'Neil, 1997; Foltz, et al., 1999; Foltz, 1996; Lemaire & Dessus, 2001; Landauer, et al., 1998; Miller, 2003; Steinhart, 2001; Wade-Stein & Kintsch, 2003), not enough research has been done on using LSA for instructional software (Lemaire & Dessus, 2001). Previous studies involved both young students and university-age students, and several different knowledge domains. An open question is how LSA can be used to improve the learning of university-age, computer science students. This section offers three characteristics that distinguish this research from existing research involving the use of LSA to analyse expository writing texts and reports on a pilot study to determine the feasibility of using LSA to mark students' short essay answers to exam questions.

### 4.1 Focuses of the experiment

This subsection describes three facets of the experiment that involve under-researched areas, in the cases of the domain and the type of assessment, and an unsolved research question in the case of the appropriate dimension reduction value for small corpora.

The study involves essays written by computer science (CS) students. CS, being a technical domain, has a limited, specialist vocabulary. Thus, essays written for CS exams are thought to have a more restricted terminology than do the expository writing texts usually analysed by LSA researchers. Nevertheless, the essays are written in English using a mixture of technical terms and general terms. Will LSA produce valid results?

Accuracy is paramount in summative assessment. Whereas formative assessment can be general and informative, summative assessment requires a high degree of precision. Can LSA produce results with a high degree of correlation with human markers?

The consensus among LSA researchers, who customarily use very large corpora, is that the number of dimensions that produces the best result is about 300. But because this study involved just 17 graded samples, the number of reduced dimensions has to be less than 17. Can LSA work with many fewer dimensions than 300? A broader question is whether LSA can work with a small corpus in a restricted domain.

### 4.2 The Data

The data for this experiment consisted of answers from six students to three questions in a single electronic exam held at the Open University in April 2002. The answers are free-form short essays. The training corpus for each question comprised 16 documents consisting of student answers to the same question and a specimen solution. Table 1 gives the average size (in words) of both the student answers graded by LSA and the corpus essays.

| | Question A | Question B | Question C |
|---|---|---|---|
| Corpus documents | 112 | 35 | 131 |
| Student answers | 108 | 31 | 88 |

Table 1: Average document size

The corpus training documents had been marked previously by three trained human markers. The average marks were assigned to each corpus document. To provide a standard on which to judge the LSA results, each of the answers from the six students was marked by three human markers and awarded the average mark.

### 4.3 The LSA Method

The following steps were taken three times, once for each question on the exam.

- Determine the words, or terms, in the corpus documents after removing punctuation and stop words. (No attempt has yet been made to deal with synonyms or word forms, such as plurals, via stemming.)
- Construct a t x d term frequency matrix M, where t is the number of terms in the corpus and d is the number of documents – 17 in this experiment. Each entry $tf_{ij}$ is the number of times term i appears in document j.
- Weight each entry $tf_{ij}$ in M using the simple weighting scheme: $1 + \log(tf_{ij})$.
- Perform singular value decomposition of the weighted term frequency matrix resulting in $M_{weighted} = TSD^T$.
- Choose an optimum dimension, k, to reduce $M_{weighted.}$ (see the next subsection for details)
- Compute $B = SD^T$ - the reduced weighted frequency document
- Construct a vector, *a*, of weighted term frequencies in a student-answer document.
- Compute the reduced student-answer vector $a' = aTS^T$
- Determine the corpus document that best matches the student-answer by comparing *a'* with the column vectors in B.
- Award the student-answer the mark associated with the most similar corpus document using the cosine similarity measure.

### 4.4 Determining the optimum dimension reduction (k)

- This experiment reduced the SVD matrices using k = 2 .. number of corpus documents – 1, or k = 2 .. 16. For each value of k, the LSA method produced a mark for each student-answer.
- The experiment compared the six LSA marks for the student-answers with the corresponding average human mark using Euclidean distance.
- The experiment revealed that, for this corpus, k = about 10 gave the best matches across the three questions.

### 4.5 Results

The four graphs below show the results obtained.

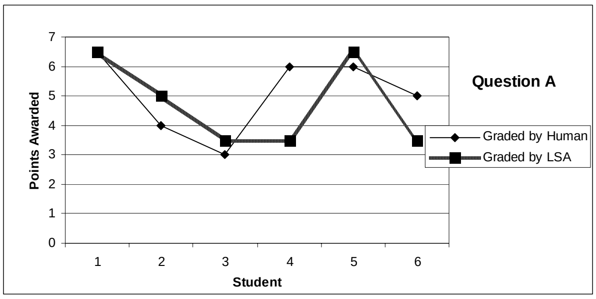

Figure 1: LSA marks for question A

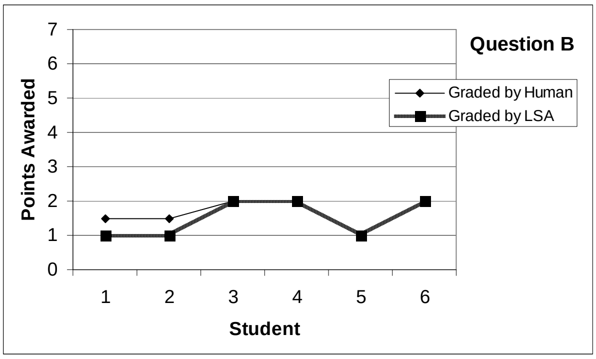

Figure 2: LSA marks for question B

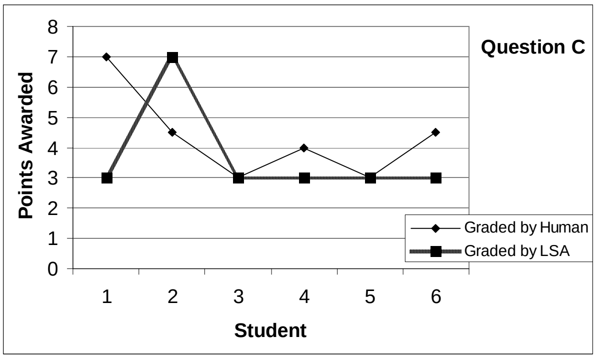

Figure 3: LSA marks for question C

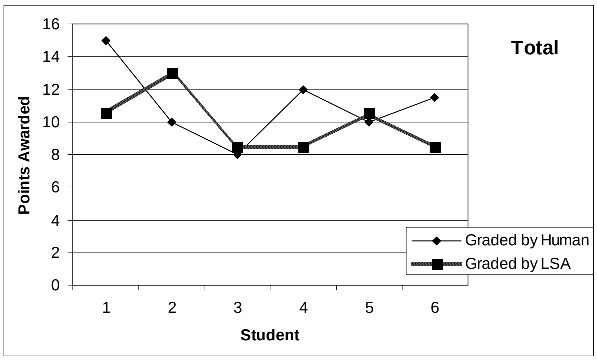

Figure 4: LSA marks for total

### 4.6 Discussion

This experiment investigated the feasibility of using LSA to assess short essay answers. The results shown in Figures 1 – 3 suggest that LSA-marked answers were similar to human-marked answers in 83% (15 of 18) of the answers tested.

LSA seemed to work well on five of the six student-answers for Question A, all the answers for Question B, and four of the six answers for Question C. For the three clearly incorrect answers, LSA gave a higher score than did the human markers for the answer to question A and one higher mark and one lower mark than did the human markers for the answers to question C.

To quantify these visual impressions, the study used the Spearman's rho statistical test for each of the three questions. Only one of the three questions shows a statistical correlation between LSA and human marks: question B shows a statistical correlation significant at the 95% level.

These results, while unacceptable for a real-world application, are encouraging given the extremely small corpus size of only 17 documents, or about 2,000 words for questions A and C and about 600 words for question B. This pilot study solidified our understanding of how to use LSA, the importance of a large corpus, and how to approach further research to improve the results and increase the applicability of the results of this pilot study.

## 5 A roadmap for further research

### 5.1 The corpus

LSA results depend on both corpus size and corpus content.

#### 5.1.1 Corpus size

Existing LSA research stresses the need for a large corpus. The corpora for the pilot study described in this paper were very small. In addition, the documents are too few in number to be representative of the student population. An ideal corpus would provide documents that give a spread of marks across the mark range and a variety of answers for each mark. Future studies will use a larger corpus.

#### 5.1.2 Corpus content

Wiemer-Hastings, et. al (1999) report that size is not the only important characteristic of the corpus. Not surprisingly, the composition of the corpus effects the results of essay grading by LSA. In addition to specific documents directly related to their essay questions, Wiemer-Hastings, et. al used more general documents. They found the best composition to be about 40% general documents and 60% specific documents.

The corpora used for this pilot study comprised only specific documents - the human marked short essays. Future work will involve adding sections of text books to enlarge and enrich the corpus with more general documents.

### 5.2 Weighting function

The pilot study used local weighting - the most basic form of term weighting. Local weighting is defined as $tf_{ij}$ (the number of times term i is found in document j) dampened by the log function: local weighting = 1 + log ($tf_{ij}$ ). This dampening reflects the fact that a term that appears in a document x times more frequently than another term is not x times more important.

The study selected this simple weighting function to provide a basis on which to compare more sophisticated functions in future work. Many variations of weighting functions exist; two are described next.

#### 5.2.1 Log-entropy

Dumais (1991) recommended using log-entropy weighting, which is local weighting times global weighting. Global weighting is defined as 1 – the entropy or noise. Global weighting attempts to quantify the fact that a term appearing in many documents is less important than a term appearing in fewer documents.

The log-entropy term weight for term i in doc j =

$$\log\left(1+tf_{ij}\right)*\left[1-\frac{\sum\frac{tf_{ij}}{gf_i}*\log\frac{tf_{ij}}{gf_i}}{\log\left(numdocs\right)}\right]$$

where

$tf_{ij}$ – term frequency – the frequency of term i in document j

$gf_i$ – global frequency – the total number of times term i occurs in the whole collection

#### 5.2.2 tfidf

Sebastiani (2002) claims the most common weighting is tfidf, or term frequency inverse document frequency.

$$tfidf\left(t_k,d_j\right)=\#\left(t_k,d_j\right)*\log\frac{|Tr|}{\#Tr\left(t_k\right)}$$

where #( $t_k$, $d_j$ ) denotes the number of times $t_k$ occurs in $d_j$

$\#Tr(t_k)$ denotes the document frequency of term $t_k$, that is, the number of documents in *Tr* in which $t_k$ occurs.

Future studies will examine the effects of applying various term weighting functions.

### 5.3 Similarity measures

The pilot study used two different similarity measures. It used the cosine measure to compare the test document with the corpus documents. It used Euclidean distance to choose k, the number of reduced dimensions that produced the best results overall. Other measures exist and will be tried in future studies.

Ljungstrand and Johansson (1998) define the following similarity measures:

Inner product (dot) measure:

$$M(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} x_i y_i$$

Cosine measure:

$$M(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}}$$

Manhattan distance measure:

$$M(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Euclidean distance measure (2-norm):

$$M(\mathbf{X},\mathbf{Y}) = \sqrt{\sum_{i=1}^{n} \left( x_i - y_i \right)^2}$$

$m$-norm measure:

$$M(\mathbf{X},\mathbf{Y}) = \left( \sum_{i=1}^{n} \left( x_i - y_i \right)^m \right)^{\frac{1}{m}},\ m \in \boldsymbol{N}$$

Where $\mathbf{X} = (x_1,x_2,...,x_n)$ and $\mathbf{Y} = (y_1,y_2,...,y_n)$ are two $n$-dimensional vectors.

Figure 5. Similarity Measures

### 5.4 Corpus pre-processing

Removing stop words is one type of pre-processing performed for this study. Explicitly adding synonym knowledge and stemming are two additional ways of preparing the corpus that future research will consider. Stemming involves conflating word forms to a common string, e.g., *write, writing, writes, written, writer* would be represented in the corpus as *writ.*

### 5.5 Dimension reduction

Choosing the appropriate dimension, k, for reducing the matrices in LSA is a well known open issue. The current consensus is that k should be about 300. No theory yet exists to suggest the appropriate value for k. Currently, researchers determine k by empirically testing various values of k and selecting the best one. The only heuristic says that k << min(terms, documents). An interesting result from the study reported in this paper is that even though k had to be less than 17, the number of documents in our corpora and thus much less than the recommended value of 300, LSA produced statistically significant results for one of the three questions tested.

Future studies will continue to investigate the relationship among k, the size of the corpus, the number of documents in the corpus, and the type of documents in the corpus.

## 6 Summary

This paper introduced and explained LSA and how it can be used to provide e-assessment by both formative and summative assessment. It provided examples of existing research that uses LSA for e-assessment. It reported the results of a pilot study to determine the feasibility of using LSA to assess automatically essays written in the domain of computer science. Although just one of the three essay questions tested showed that LSA marks were statistically correlated to the average of three human marks, the results are promising because the experiment used very small corpora.

Future studies will attempt to improve the results of LSA by increasing the size of the corpora, improving the content of the corpora, experimenting with different weighting functions and similarity measures, pre-processing the corpus, and using various values of k for dimension reduction.

## 7 Acknowledgements

The work reported in this study was partially supported by the European Community under the Innovation Society Technologies (IST) programme of the 6th Framework Programme for RTD - project ELeGI, contract IST-002205. This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

## 8 References

Chung, G., & O'Neil, G. (1997). *Methodological approaches to online scoring of essays* (Center for the Study of Evaluation, CRESST No. 461). Los Angeles.

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*(6), 391-407.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavioral Research Methods, Instruments & Computers, 23*(2), 229-236.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197-202.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99.*

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., et al. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *ACM,* pp. 465-480.

Furnas, G. W., Gomez, L. M., Landauer, T. K., & Dumais, S. T. (1982). Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 251-253). ACM.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research, 24*(3), 305-320.

Ljungstrand, P., & Johansson, H. (1998, May). *Intranet indexing using semantic document clustering*. Retrieved 5/4/2004, from http://www.handels.gu.se/epc/archive/00002294/01/ljungstrand.IA7400.pdf.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press.

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research, 28.*

Sebastiani, F. (2002, March). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Steinhart, D. J. (2001). *Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis.* Unpublished doctoral dissertation, University of Colorado, Boulder, Department of Psychology.

Wade-Stein, D., & Kintsch, E. (2003). *Summary Street: Interactive computer support for writing* (Tech Report from the Institute for Cognitive Science). University of Colorado, USA.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education.* Amsterdam: IOS Press.