

Seeing the Whole Picture: Evaluating Automated Assessment Systems

Debra Trusso Haley, Pete Thomas, Anne De Roeck, Marian Petre
The Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes MK7 6AA UK
D.T.Haley, P.G.Thomas, M. Petre, A.DeRoeck at open.ac.uk

Abstract: This paper argues that automated assessment systems can be useful for both students and educators provided that its results correspond well with human markers. Thus, evaluating such a system is crucial. We present an evaluation framework and show why it can be useful for both producers and consumers of automated assessment. The framework builds on previous work to analyse Latent Semantic Analysis- (LSA) based systems, a particular type of automated assessment, that produced a research taxonomy that could help developers publish their results in a format that is comprehensive, relatively compact, and useful to other researchers. The paper contends that, in order to see a complete picture of an automated assessment system, certain pieces must be emphasised. It presents the framework as a jigsaw puzzle whose pieces join together to form the whole picture and provides an example of the utility of the framework by presenting some empirical results from our assessment system that marks questions about html. Finally, the paper suggests that the framework is not limited to LSA-based systems. With slight modifications, it can be applied to any automated assessment system.

Keywords: automated assessment systems, computer aided assessment, CAA, Latent Semantic Systems, LSA systems; teaching programming

1. Introduction

1.1. Arguments for and against using an automated assessment system

Assessment is an important component of teaching programmers. Researchers (Berglund, 1999; Daniels, Berglund, Pears, & Fincher, 2004) report that assessment can have a strong effect on student learning. Students learn best by frequent assessment with rapid feedback. Unfortunately, assessment can be an onerous task for educators. It takes time both to create the assessments and to mark them. Computers can reduce the time humans spend marking assessments. The educators can then use their time for more creative work. Educational institutions hope to save time, and therefore, money by using computerised marking systems.

In addition to the possible time and cost savings, a computer offers some advantages over humans. Human markers may mark differently as they become fatigued as well as being affected by the order of marking. For example, if a marker first encounters a brilliant answer, the experience could cause the marker to be harsher for the remainder of the answers. Even the most scrupulous people might show bias based on personal feelings towards a student. While they may successfully avoid awarding better marks to their favourite students, they may mark non-favoured students more highly than they deserve in an attempt to be unbiased. Automatic markers can be an improvement over human markers because their results are reliable and repeatable. They do not get tired, they do not show bias based on personal feelings towards students, their results will be the same without regard to the order in which the answers are presented, and they are able to return results much faster than humans.

The major objection to using automated assessment is concern over its accuracy. Not only is there no agreed-upon *level* of acceptable accuracy, there is no agreed-upon *method* by which to measure the accuracy of automated assessment system systems. Evaluation of the marking systems is a crucial topic because they will not be used if people do not have faith in their accuracy. We contend that an acceptable accuracy level would match the rate at which human markers correspond with each other.

Another objection is that automatic marking takes away the human touch. We offer the suggestion that if an educator uses automatic marking, the time saved can be devoted to more personal contact with students. In addition, we would not entirely replace human markers with a computer. Our university uses multiple markers for high-stakes exams. A panel of experienced markers then moderates the marks where the humans don't agree. An automatic assessment system could take the place of one of the human markers. By using a human and a computer to mark the same questions, educators can benefit from double-checking the computer with the human and vice versa.

1.2. Some existing assessment systems

Various automated assessment systems have been created to save time by automating marking. CourseMarker is an automated assessment tool for marking programs (<http://www.cs.nott.ac.uk/~ceilidh/>). Other automated assessment systems mark essays or short answers. For example, see (Burstein, Chodorow, & Leacock, 2003) for an assessment system that grades general knowledge essays and (Wiemer-Hastings, Graesser, & Harter, 1998) for a tutoring system that evaluates answers in the domain of computer science.

As part of our work to improve the learning of programming and computing in general, we research automated assessment systems. We have developed a tool (Thomas, Waugh, & Smith, 2005) that is part of an online system to mark diagrams produced by students in a database course. We are developing EMMA (ExaM Marking Assistant) a Latent Semantic Analysis-(LSA) based automated assessment system (D. Haley, Thomas, De Roeck, & Petre, 2007) to mark short answers about html and other areas in computer science. LSA is a statistical natural language processing technique for analysing the meaning of text. We chose LSA because it has been used successfully in the past to mark general knowledge essays (Landauer, Foltz, & Laham, 1998) and shows promise in our area of short answers in the domain of computer science. This paper does not offer an LSA tutorial. Readers desiring a basic introduction to LSA should consult the references section (Landauer et al., 1998). We discuss LSA only as necessary to justify the need for our taxonomy and evaluation framework.

Our work with EMMA has highlighted a significant challenge – the developer must choose many options that are intrinsic to the success of any LSA-based marking system. A review of the literature (D. T. Haley, Thomas, De Roeck, & Petre, 2005) revealed that although many researchers have reported work based on LSA, it is difficult to get a full picture of these systems. Some of the missing information includes type of training material and examples of questions being marked as well as fundamental LSA options, e.g., weighting function and number of dimensions in the reduced matrix.

1.3. Central theme of the paper

The aim of this paper is to offer our two-part framework for automated assessment systems and to explain why it is necessary. It is based on a research taxonomy (D. T. Haley et al., 2005) we developed to compare Latent Semantic Analysis (LSA) based educational applications. The framework can be of value to both producers and consumers of automated assessment systems.

Producers are researchers and developers who design and build assessment systems. They can benefit from the framework because it provides a relatively compact yet complete description of relevant information about their systems. If producers of automated assessment systems use the framework, they can contribute to the improvement of the state-of-the-art by adding to a collection of comparable data.

Consumers are organisations, such as universities, that wish to use an automated assessment system. These consumers are, or should be, particularly interested in two areas. The first and most important area is the accuracy of the results. But what does accuracy mean and how do we measure it? We believe that an automated assessment system is *good enough* if its marks compare to human markers as well as human markers compare with each other. We have discussed various ways of measuring accuracy in previous work (D. Haley et al., 2007). Second, consumers should be interested in the amount of human effort required to use the assessment system. Most natural language processing assessment systems, including those based on LSA, require a large amount of training data. Although the system might save time for markers, it may take too much time to prepare the system for deployment (for example, to train the system for a specific data set) .

It is difficult to compare automatic assessment systems because no uniform procedure exists for reporting results. This paper attempts to fill that gap by proposing a framework for reporting on and evaluating automatic assessment tools.

2 The framework

The first part of the framework for describing an automated assessment system can be visualised as the jigsaw puzzle in Figure 1. Figure 2 shows the second part of the framework – the evaluation of the system. We contend that all the pieces of this puzzle must be present for a reviewer to see the whole picture.

The important categories of information for specifying an automated assessment system are the items assessed, the training data, and the algorithm-specific technical details. The general type of question (e.g., essay and multiple choice) is crucial for indicating the power of a system. The granularity of the marking scale provides important information about the accuracy – it is usually easier for two markers to agree when they grade a 3 point question than one worth 100 points. The number of items assessed provides some idea of the generalise-ability and validity of the results. Both the number of unique questions and the number of examples of each question contribute to the understanding of the value of the results. The second category comprises the technical details of the algorithm used. Haley, et al (2005) discuss why these options are of interest to producers of an LSA-based automated assessment system. The central piece of Figure 1 shows LSA-specific options, but these would be changed if the automated assessment system is based on a different method. The data used to train the system is another crucial category. Both the type and amount of text help to indicate the amount of human effort needed to gather this essential element of automated assessment systems. Some systems (LSA for one (D. Haley et al., 2007)) need two types of training data – general text about the topic being marked and specific previously marked answers for calibration. Researchers should give details about both these types of training data.

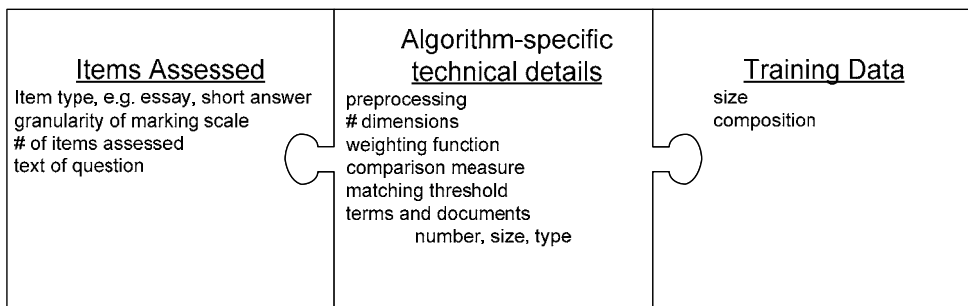


Figure 1. First part of framework: comparing automated assessment systems

Anyone interested in developing or using an automated assessment system will be interested in its evaluation. The accuracy of the marks is of primary importance. An automated assessment system exhibiting poor agreement with human markers is of little value. Our previous work (D. T. Haley et al., 2005) showed that different researchers report their results using different methods. Ideally, all researchers would use the same method for easily comparable results. If researchers fail to reach a consensus on what information should be reported, they should at least clearly specify how they determined the accuracy of their results. The other two pieces of the evaluation picture are usability and effectiveness. These pieces are of interest to consumers wanting to choose among deployed systems.

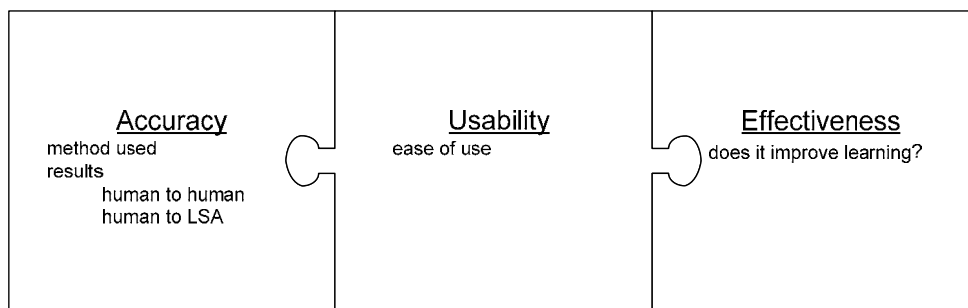


Figure 2. Second part of framework: evaluating automated assessment systems

3 Research taxonomy for LSA-based automated assessment systems

This section summarises a research taxonomy developed in (D. T. Haley et al., 2005). It was the result of an in-depth, systematic review of the literature concerning Latent Semantic Analysis (LSA) research in the domain of educational applications. The taxonomy was designed to present and summarise the key points from a representative sample of the literature.

The taxonomy highlighted the fact that others were having difficulty matching the results reported by the original LSA researchers (Landauer & Dumais, 1997). We found a lot of ambiguity in various critical implementation details (e.g. weighting function used) as well as unreported

details. We speculated that the conflicting or unavailable information explains at least some of the inability to match the success of the original researchers.

The next subsections discuss the rationale for choosing certain articles over others and the meaning of the headings in the taxonomy.

3.1. Method for choosing articles

The purpose of the taxonomy was to summarise and highlight important details from the LSA literature. Because the literature is extensive and our interest is in the assessment of essays and related artefacts, the taxonomy includes only those LSA research efforts that overlap with educational applications. The literature review found 150 articles of interest to researchers in the field of LSA-based educational applications. In order to limit this collection to a more reasonable sample, we constructed a citer – citee matrix of articles. That is, each cell entry (i, j) was non blank if article i cited article j . The articles ranged in date from perhaps the first LSA published article (Furnas et al., 1988), to one published in May 2005 (Perez et al., 2005). We found the twenty most-cited articles and placed them, along with the remaining 130 articles, in the categories shown in Table 1.

Type of Article	Number in Lit Review	Number in Taxonomy
most cited	20	13
LSA and ed. applications	43	15
LSA but not ed. apps.	13	0
LSI	11	0
theoretical / mathematical	11	0
reviews / summaries	11	0
ed. apps. but not LSA	41	0
Total	150	28

Table 1. Categories of articles in the literature review and those that were selected for the taxonomy

We chose the twenty most-cited articles for the taxonomy. Some of these most-cited articles were early works explaining the basic theory of Latent Semantic Indexing (LSI).¹ Although not strictly in our scope of the intersection of LSA and educational applications, we included some of these articles because of their seminal nature. Next, we added articles from the category that

¹ Researchers trying to improve information retrieval produced the LSI theory. Later, they found that LSI could be useful to analyse text and created the term LSA to describe LSI when used for this additional area.

combined educational applications with LSA that were of particular interest, either because of a novel domain or technique, or an important result. Finally, we decided to reject certain heavily cited articles because they presented no new information pertinent to the taxonomy. This left us with 28 articles in the taxonomy.

3.2. The taxonomy categories

The taxonomy organises the articles involving LSA and educational applications research into three main categories: an *Overview*, *Technical Details*, and *Evaluation*. Figures 3, 4, and 5 show the headings and sub-headings. Most of the headings are self-explanatory; some clarifications are noted in the figures.

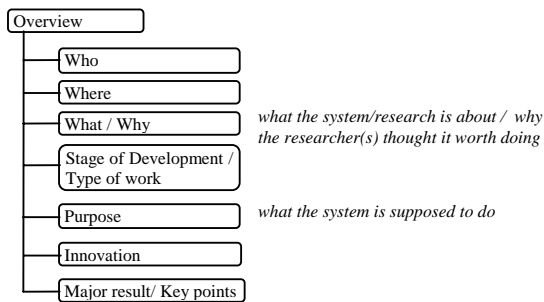


Figure 3. Category A: Overview

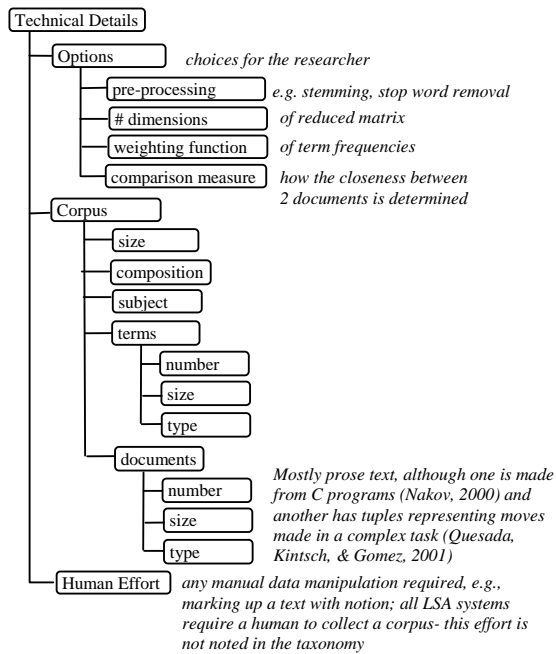


Figure 4. Category B: Technical Details

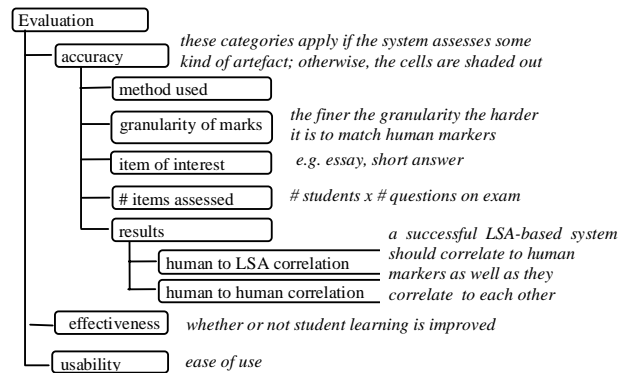


Figure 5. Category C: Evaluation

Appendix A presents the taxonomy. When looking at it, the reader should keep a few points in mind. First, the taxonomy is three pages wide by three pages high. Pages 1-3 cover the overview for all of the articles in the taxonomy. Pages 4-6 list the technical details. Pages 7-9 give the evaluation information. Second, each line presents the data relating to one study. However, one article can report on several studies. In this case, several lines are used for a single article. The cells that would otherwise contain identical information are merged. Third, the shaded cells indicate that the data item is not relevant for the article being categorised. Fourth, blank cells indicate that we were unable to locate the relevant information in the article. Fifth, the information in the cells was summarised or taken directly from the articles. Thus, the *Reference* column on the far left holds the citation for the information on the entire row.

Organising a huge amount of information in a small space is not easy. The taxonomy in the appendix is based on an elegant solution in (Price, Baecker, & Small, 1993).

4 Using the Framework for an automated assessment system

Our framework for evaluating an automated assessment system is a refined version of the taxonomy discussed in the previous section. The experience of creating and using the taxonomy served to crystallize our thinking about the important elements of reporting on an automated assessment system. Table 2 is an example of how the framework could be used to compare different systems in tabular form. It starts with an overview and proceeds with the pieces in the puzzles of Figures 1 and 2.

Table 2. Filling in the framework

System Name	Overview					Items Assessed					Algorithm-specific Technical Details									
	Reference	Who / Where	What / Why	Stage of Development / Type of Innovation	Major Result / Key points	Human Effort	Type of Item	Granularity of Marking Scale	# of items assessed	text of question	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Terms	Documents			
HTD07	HTD07	Hayley, Thomas, DeRoock, Peire; The Open University	assess computer science short answers for summative assessment	research prototype	amount of training data that works best: 50 marked answers for question A	gather training data, gather marked answers	short answers about html	4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML The desired appearance is very is very It is very important to read this text carefully. The desired appearance Things to do: Pack suitcase. Book taxi.	stemming, stop words	90	log entropy	cosine	none	12k words	text	45k paragraphs	1	text
EMMA					amount of training data that works best: 80 marked answers for B			4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML The desired appearance Things to do: Pack suitcase. Book taxi.		500	log entropy	cosine	none					

Reference	Training Data			Evaluation				Usability	
	Size	Composition	method used	accuracy	Human to LSA	Human to Human	Effectiveness	Usability	
HTD07	1) 45k paragraphs 2) 50	1) course texts 2) human marked answers	compared LSA marks with 5 human markers and calculated average	average % identical off by 1 off by 2 off by 3 off by 4	53 34 12 1 1 1	54 32 11 1 1	Does it improve learning? not relevant - a research prototype	How easy is it to use? not relevant - a research prototype	
	1) 45k paragraphs 2) 80	1) course texts 2) human marked answers	compared LSA marks with 5 human markers and calculated average	identical off by 1 off by 2 off by 3 off by 4	43 45 6 3 3	61 28 9 1 1	not relevant - a research prototype	not relevant - a research prototype	

Our previous work (D. T. Haley et al., 2005) highlighted the insights revealed by the taxonomy. The major conclusion was that researchers need to know all of the details to fully evaluate and compare reported results. The taxonomy contains many blank cells. This implies that much valuable information goes unreported. Research results cannot be reproduced and validated if researchers do not provide more detailed data regarding their LSA implementations.

The framework (see figures 1 and 2) is an attempt to simplify the taxonomy and make it more concise. The information reports the results of a previous study (D. Haley et al., 2007) to determine the optimum amount of training data to mark questions about html. All of the relevant information concerning that study is in the table. The assessment system is called EMMA. It was developed by Haley, et al. to assess computer science short answers for summative assessment. EMMA is a research prototype – not yet a deployed system. The innovation of the study was to determine the optimum amount of training data and found that 50 marked answers were optimum for question A and 80 marked answers were optimum for question B. Each of the questions about html was worth 4 points and we evaluated 50 student answers per question. The table contains the text of the two questions. The table gives the information relating to LSA parameters. This may not be of interest to consumers of assessment systems but is vital for other researchers wishing to replicate the findings. We used 45,000 paragraphs from course textbooks to serve as general training data. To evaluate the results of EMMA, we compared the marks given by five humans and calculated the average. We then compared EMMA's marks with each of the five humans and calculated the average. We found that EMMA worked better for question A than it did for Question B. Fifty-three percent of EMMA's marks were identical to the human marks. Thirty-four percent of the marks differed by one point, 12% differed by two points, and 1% differed by three and four points. This compares to the human average agreement, which was 54, 32, 11, 1, and 1 for the same point differences. These figures suggest that EMMA produced very similar results to what the humans did for question A. The results were not as good for question B. The table gives the relevant figures.

The previous paragraph repeats the information in the table. It is easier to use the table to compare our results with other system than it is to digest the text in the previous paragraph. The table gives all of the information specified in the framework in a reasonably concise form.

5 Conclusions

Our framework will support sharing and comparison of results of further research into LSA-based automated assessment system tools. By providing all the pieces of the puzzle, researchers show the whole picture of their systems. The publication of all relevant details will lead to improved understanding and the continued development and refinement of LSA.

Our work has involved an LSA-based system. However, the same benefits that accrue to LSA researchers by using the framework can also extend to broader automated assessment system research. The framework can be altered by replacing the LSA-specific technical details with the relevant information.

We hope that by presenting this framework, we stimulate discussion amongst automated assessment system producers and consumers. The ultimate goal is to improve computing education by improving assessment.

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.

References

- Bassu, D., & Behrens, C. (2003). Distributed LSI: Scalable concept-based information retrieval with high semantic resolution. In *Proceedings of Text Mining 2003, a workshop held in conjunction with the Third SIAM Int'l Conference on Data Mining*. San Francisco.
- Berglund, A. (1999). *Changing Study Habits - a Study of the Effects of Non-traditional Assessment Methods*. Work-in-Progress Report. Paper presented at the 6th Improving Student Learning Symposium, Brighton, UK.
- Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review* 37, 4, 573-595.
- Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion 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.
- Daniels, M., Berglund, A., Pears, A., & Fincher, S. (2004). *Five Myths of Assessment*. Paper presented at the 6th Australasian Computing Education Conference (ACE2004), Dunedin, New Zealand.
- 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., Laham, D., & Landauer, T. K. (1999). *Automated Essay Scoring: Applications to Educational Technology*. Paper presented at the ED-MEDIA '99, Seattle.
- 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. In *Proceedings of 11th annual int'l ACM SIGIR conference on Research and development in information retrieval* (pp. 465-480): ACM.
- Haley, D., Thomas, P., De Roeck, A., & Petre, M. (2007, 31 January 2007). *Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML*. Paper presented at the Proceedings of the Ninth Australasian Computing Education Conference (ACE2007), Ballarat, Victoria, Australia.
- Haley, D. T., Thomas, P., De Roeck, A., & Petre, M. (2005, 21-23 September 2005). *A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications*. Paper presented at the International Conference on Recent Advances in Natural Language Processing'05, Borovets, Bulgaria.
- Kanejiya, D., Kumar, A., & Prasad, S. (2003). *Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA*. Paper presented at the HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing.
- Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*. [Special Issue, J. Psozka, guest editor], 8(2), 87-109.
- 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.
- Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412-417).
- Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24(3), 305-320.
- Nakov, P. (2000). Latent Semantic Analysis of Textual Data. In *Proceedings of the Int'l Conference on Computer Systems and Technologies*. Sofia, Bulgaria.

- Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. In *Proceedings of the EuroConference Recent Advances in Natural Language Processing (RANLP'01)*. Tzigov Chark, Bulgaria.
- Olde, B. A., Franceschetti, D. R., Karnavat, A., & Graesser, A. C. (2002). The Right Stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 708-713). Fairfax.
- Perez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodriguez, P., & Magnini, B. (2005). *Automatic Assessment of Students' free-text Answers underpinned by the combination of a Bleu-inspired algorithm and LSA*. Paper presented at the Proceedings of the 18th Int'l FLAIRS Conference, Clearwater Beach, Florida.
- Price, B. A., Baecker, R. M., & Small, I. S. (1993). A Principled Taxonomy of Software Visualization. *Journal of Visual Languages and Computing*, 4(3), 211-266.
- Quesada, J., Kintsch, W., & Gomez, E. (2001). A computational theory of complex problem solving using the vector space model (part 1): Latent Semantic Analysis, through the path of thousands of ants. *Cognitive Research with Microworlds*, 43-84, 117-131.
- Thomas, P. G., Waugh, K., & Smith, N. (2005). *Experiments in the Automatic Marking of ER-Diagrams*. Paper presented at the Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, Monte de Caparica, Portugal.
- Wiemer-Hastings, P., Graesser, A., & Harter, D. (1998). The foundations and architecture of Autotutor. In *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 334-343). San Antonio, Texas.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education*. Amsterdam: IOS Press.
- Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for Syntax, Vectors for Semantics. In *Proceedings of the 23rd Cognitive Science Conference*.
- Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., et al. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, 25, 309-336.

Appendix A
The Latent Semantic Analysis Research Taxonomy

System Name	Refer-ence	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
Indexing not assessing essays	DDF90	Deerwester, Dumais, Furnas, Landauer, Harshman	U of Chicago, Bellcore, U of W. Ontario	explain new theory that overcomes the deficiencies of term-matching	LSI research	information retrieval	LSI: explains SVD and dimension reduction steps	for Med: for all but the two lowest levels of recall, precision of the LSI method lies well above that obtained with straight-forward term matching; no difference for CISI
	Dum91	Dumais	Bellcore	attempt better LSI results	LSI research	information retrieval	compared different weighting functions	log entropy best weighting function; stemming and phrases showed only 1-5% improvement; 40% better than raw frequency weighting
Indexing	BD095	Berry, Dumais, O'Brien	U of Tenn, Bellcore	explain new theory	LSI research	information retrieval	LSI	LSI - completely automatic indexing method using SVD, shows how to do SVD updating of new terms
	FBP94	Folz, Brit, Perfetti	New Mexico State University, Slippery Rock U, U of Pittsburgh	matching summaries to text read, determine if LSA can work as well as coding propositions	LSA research	text comprehension to evaluate a reader's situation model	matching summaries to text read, analyses knowledge structures of subjects and compares them to those generated by LSA	representation generated by LSA is sufficiently similar to the readers' situation model to be able to characterize the quality of their essays
	FKL98	Folz, Kintsch, Landauer		measure text coherence	LSA research		using LSA to measure text coherence	LSA needs a corpus of at least 200 documents; online encyclopedia articles can be added
	LD97	Landauer, Dumais	U of Colorado, BellCore	explain new theory	LSA research			LSA could be a model of human knowledge acquisition
	LLR97	Landauer, Laham, Rehder, Schreiner	U of Colorado	compared essays scores given by readers and LSA, to determine importance of word order	LSA theory	grading essays	investigating the importance of word order; combined quality (cosine) and quantity (vector length)	LSA predicted scores as well as human graders; separating tech and non-technical words made no improvement
	RSW98	Rehder, Shreiner, Wolfe, Laham, Landauer, Kintsch	U of Colorado	explore certain technical issues	LSA research	grading essays	investigated technical vocabulary, essay length, optimal measure of semantic relatedness, and directionality of knowledge in the high dimensional	nothing to be gained by separating essay into tech and non tech terms cosine and length of essay vector are best predictors of mark
	WSR98	Wolfe, Shreiner, Rehder, Laham, Foltz, Kintsch, Landauer	U of Colorado, New Mexico State Univ	compared essay scores after reading one of 4 texts	LSA research	select appropriate text	using LSA to select appropriate text	LSA can measure prior knowledge to select appropriate texts
Intelligent Essay Assessor (IEA) http://psych.msu.edu/essay	FLL99	Folz, Landauer, Laham	New Mexico State University, Knowledge Analysis Technologies, U of Colorado	reports on various studies using LSA for automated essay scoring	deployed application for formative assessment	practice essay writing		Over many diverse topics, the IEA scores agreed with human experts as accurately as expert scores agreed with each other.

System Name	Reference	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
Summary Street	KSS00	Kintsch, Steinhart, Stahl, LSA Research Group, Matthews, Lamb	U of Colorado, Platt Middle School http://www.k-a-t.com/cu.shtml	helps students summarize essays to improve reading comprehension skills	deployed application for formative assessment	provide feedback on length, topics covered, redundancy, relevance	graphical interface, optimal sequencing of feedback	students produced better summaries and spent more time on task with Summary Street
Summary Street	Ste01	Steinhart	U of Colorado http://www.k-a-t.com/cu.shtml	helps students summarize essays to improve reading comprehension skills	deployed application for formative assessment	provide feedback on length, topics covered, redundancy, relevance	graphical interface, optimal sequencing of feedback	the more difficult the text, the better was the result of using Summary Street, feedback doubled time on task
	Lan02b	Landauer	U of Colorado	explaining LSA		LSA general research		LSA works by solving a system of simultaneous equations
AutoTutor	WWG99	Wiemer-Hastings, P., Wiemer-Hastings, K, Graesser, A.	U of Memphis	test theory that LSA can facilitate more natural tutorial dialogue in an intelligent tutoring system (ITS)	deployed application for formative assessment	assess short answers given to intelligent Tutoring System	tested size and composition of corpus for best LSA results	LSA works best when specific texts comprise at least 1/2 of the corpus and the rest is subject related; works best on essays > 200 words
	Wie00	Wiemer-Hastings	U of Memphis	determine effectiveness of adding syntactic info to LSA	LSA research	assess short answers given to ITS	added syntactic info to LSA	adding syntax decreased the effectiveness of LSA - as compared to Wie99 study
Select-a-Kibitzer	WG00	Wiemer-Hastings, Graesser	U of Memphis	give meaningful feed back on essays using agents	deployed application for formative assessment	assess short answers given to ITS	investigated types of corpora for best results	best corpus is specific enough to allow subtle semantic distinctions within the domain, but general enough that moderate variations in terminology won't be lost
SLSA - Structured LSA	WZ01	Wiemer-Hastings, Zipitria	U of Edinburgh	evaluate student answers for use in ITS	LSA research	assess short answers given to ITS	combines rule-based syntactic processing with LSA - adds part of speech	adding structure-derived information improves performance of LSA; LSA does worse on texts < 200 words
	Nak00b	Nakov	Sofia University	explore uses of LSA in textual research	LSA research		uses correlation matrix to display results; analysis of C programs	
	NPM01	Nakov, Popova, Mateev	Sofia University	evaluate weighting function for text categorisation	LSA research	analyse English literature texts	compared 2 local weighting times 6 global weighting methods	log entropy works better than classical entropy

System Name	Reference	Who	Where	What / Why	Stage of Development/ Type of work	Purpose	Innovation	Major Result / Key points
	FKM01	Franceschetti, Karnavat, Marineau, et al	U of Memphis	constructing different types of physics corpora to evaluate best type for an ITS	LSA research for formative assessment	intelligent tutoring	used 5 different corpora to compare vector lengths of words	carefully constructed smaller corpus may provide more accurate representation of fundamental physical concepts than much larger one
	OFK02	Olde, Franceschetti, Karnavat, et al	U of Memphis, CHI Systems	evaluate corpora with different specificities for use in ITS	LSA research for formative assessment	intelligent tutoring	used 5 different corpora to compare essay grades	sanitizing the corpus provides no advantage
Apex	LD01	Lemaire, Dessus	U of Grenoble-II	web-based learning system, automatic marking with feedback	deployed application for formative assessment	provide feedback on topic, outline and coherence		LSA is a promising method to grade essays
	QKG01a	Quesada, Kintsch, Gomez	U of Colorado, U of Grenada	investigate complex problem solving using LSA	CPS and LSA research		represent actions taken in a Microworld as tuples for LSA	LSA is a promising tool for representing actions in Microworlds.
Distributed LSI	BB03	Bassu, Behrens	Telcordia	improve LSI by addressing scalability problem	LSI research	information retrieval	subdivide corpus into several homogeneous subcollections	a divide-and-conquer approach to IR not only tackles its scalability problems but actually increases the quality of returned documents
SLSA	KKP03	Kanejiya, Kumar, Prasad	Indian Institute of Technology	evaluate student answers in an ITS	LSA research	intelligent tutoring	augment each word with POS tag of preceding word, used 2 unusual measures for evaluation: MAD and Correct vs False evaluation	SLSA has limited improvement over LSA
indexing not assessing essays	NVA03	Nakov, Valchanova, Angelova	U of Cal, Berkeley, Bulgarian Academy of Sciences	investigating the most effective meaning of "word"	LSA research	text categorisation	compared various methods of term weighting with NLP pre-processing	linguistic pre-processing (stemming, POS annotation, etc) does not substantially improve LSA, proper term weighting makes more difference
	THD04	Thomas, Haley, DeRoock, Petre	The Open University	assess computer science essays	LSA research for summative assessment	assess essays	used a very small, very specific corpus necessitating a small # of dimensions	LSA works ok when the granularity is coarse; need to try a larger corpus
Atenea	PGS05	Perez, Gliozzo, Strapparava, Alfonseca, Rodriguez, Magnini	U de Madrid; Istituto per la Ricerca Scientifica e Tecnologica	web-based system to assess free-text answers	LSA + ERB research		combine LSA with a BLEU-inspired algorithm; ie combines syntax and semantics	achieves state-of-the-art correlations to the teachers' scores while keeping the language-independence and without requiring any domain specific knowledge

Reference	Options				Training Data						Human Effort			
	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Terms		Documents				
								Number	Size	Type		Number	Size	Type
DDF90	remove 439 stop words (from SMART)	100		cosine		MED	medical abstracts	5,823	words		1,033	average 50	title and abstract	
Dum91	remove 439 stop words (from SMART)	60, 100	log entropy	cosine		CISI MED, CISI, CRAN, TIME, ADI	information science abstracts various (described in paper)	5,135 374 - 5831	words words		1,460	avg 45 words	title and abstract	
BD095	none	70-100	log entropy	cosine										none
FBP94		100		cosine	27.8 K	21 articles about the Panama Canal; 8 encyclopedia articles, excerpts from 2 books	Panama Canal	4829	word	prose text	607			
LD97		100		cosine		21 articles on the the heart	heart	2,781	words	prose text				
LLR97	remove 439 stop words	300	ln(1+freq)/entropy	cosine	4.6M	Grolier's Academic American Encyclopedia		60.7k	word	prose	30.4k	average 151	words	
RSW98	no stop words	94		cosine vector length		27 articles from Grolier's Academic Amer. Encyclopedia	heart anatomy	3034	word	prose	830	sentence	words	
WSR98		1500				textbook	psychology	19,153	words	prose	4,904	paragraphs	words	separated essays into technical and non technical created subsections of essays
FLL99				cosine	17,880	36 encyclopedia articles	heart anatomy							
						a portion of the textbook	psycholinguistics							
							standardised test - opinion essays							
							standardised test - argument essays							
							diverse							

Reference	Options				Training Data				Documents				Human Effort		
	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Number	Terms		Number	Size		Type	
									Size	Type					
KSS00	correct spelling			cosine		specialized texts	heart and lung	17,688	1 word	prose text	830		prose text	no pregraded summaries but mark up text into topics to appear in summaries	
Ste01	correct spelling		cosine	cosine	general knowledge space	sources of energy	heart and lung	46,951	1 word	prose text				prose text	
															specialized texts
															Meso-American history
Lan02		300		cosine		specialized texts	Meso-American history								
WWG99		200	log entropy	cosine	2.3 MB	2 complete computer literacy textbooks, ten articles on each of the tutoring topics, entire curriculum script including expected good answers	computer literacy								collect good and bad answers
Wie00	yes, see human effort			cosine			computer literacy		1 tuple	subject - verb - object		1 tuple	subject - verb - object	segmented sentences into subject, verb, object tuples; resolved anaphora; resolved ambiguities with "and" and "or"	
WGO0														researcher's task to find or create appropriate texts to serve as the corpus and comparison texts	
WZ01	removed 440 stop words			cosine	2.3 MB	same as WWG99	computer literacy							segmented sentences into subject, verb, object tuples; resolved anaphora and ambiguities with "and" and "or"	
Nak00b	removed 938 stop words	30	log and or entropy			religious texts C programs	religion	20,433			196		C code		
NPM01	removed stop words and those occurring only once	15	6 different	dot product / cosine	974 K	Huckleberry Finn and Adventures of Sherlock Holmes		5534	words	prose	487	2 KB	prose		

Reference	Options					Training Data							Human Effort	
	Pre-processing	# dimensions	Weighting function	Comparison measure	Size	Composition	Subject	Terms		Documents				
								Number	Size	Type	Number	Size		Type
FKM01		300		cosine		physics text book and other science text books	physics				paraph	paraph	prepare specialised corpora	
OFK02		300		cosine		physics text book + related to curriculum script	physics	from 1,564 to 6,536	word	prose	paraph	paraph	sanitize corpora; write "expectations" for each answer	
LD01					290K + size of course text	3 French novels plus course text	sociology of education						no pre-graded essays; mark up text into topics and notions	
QKG01a						tuples representing actions in a Microworld	complex problem solving	75565	1	tuple		1 trial		
BB03							various						create a classification scheme for LSI vector spaces	
KKP03	removed stop words		log entropy	cosine	2.3M	used Auto tutor corpus	computer literacy	9,194	word	word - part of speech tags	paraph	paraph	part of speech tagging	
NVA03	removed 442 stop words, stemming, POS	0, 10, 220, 40	various			Bulgarian	various - see paper for details							
THD04	none	10	log (no global weighting)	cosine	< 2,000	human marked answers to the essays	computer literacy				17	1 paraph	prose	none
PGS05			tf-idf			10 different corpora: student answers + text from popular computer magazines								

Reference	Accuracy						Effectiveness	Usability
	Method used	Granularity of marks	Item of Interest	Number items assessed	Results			
					Human to LSA correlation	Human to Human correlation		
DDF90	evaluate using recall and precision		queries	30				
Dum91	evaluate using recall and precision		queries	35				
BD095	evaluate using recall and precision							
FBP94	compare against human graders	100	essay	24	0.68	.367 to .768		
	compared sentences with cosine measure							
LD97			TOEFL - multiple choice test	80	LSA: 64.4%; students: 64.5%			
LLR97	compare against human graders gold standard - a short text written by an expert compare against human graders	5	short essay - 250 words	94	0.77	0.77		
				94	0.72			
RSW98	compare with 1 or more target texts		short essay - 250 words	273	0.64	0.65		
				106				
WSR98	compared with 4 texts of increasing difficulty and specificity		essay of about 250 words	106	0.63	0.77		
FLL99	holistic - compare with graded essays		essays		0.8	0.73	average grade 85; after revisions, average grade 92	survey showed 98% of students would definitely or probably use system
				695	0.86	0.86		
				668	0.86	0.87		
				1,205	0.701	0.707		

Reference	Accuracy				Item of Interest	Number items assessed	Results		Effectiveness	Usability
	Method used	Granularity of marks	Human to LSA correlation	Human to Human correlation						
KSS00	compare with teacher - provided topic list	10		summary of essay				no sig difference	in classroom 1997-1999: students like immediate feed back	
		10	0.64	summary of essay	50	0.69	scores of those using SS for difficult texts significantly higher than those not using SS	in classroom 1997-1999: students like immediate feed back		
Ste01		5		summary of essay	108		scores of those using SS for difficult texts significantly higher than those not using SS			
		5		essay	52					
		10		essay	52					
Lan02	holistic, Pearson product-moment correlation coefficient	5 or 10 points	0.81	essay	3,500	0.83				
WWG99	compare against pre-graded answers for completeness and compatibility	2: threshold of .55	0.49	short answers average length is 16 words	192	0.51				
Wle00	compared tuples in student answer with tuples in expected answer		.18, .24, and .4							
WG00										
WZ01	evaluate two texts using cosine									
Nak00b	created correlation matrices									
NPM01	defined precision as ration of chunks from same text to num of chunks at a level									

Reference	Accuracy						Effectiveness	Usability
	Method used	Granularity of marks	Item of Interest	Number items assessed	Results			
					Human to LSA correlation	Human to Human correlation		
FKM01	compared vector lengths of words for 5 different corpora							
OFK02	compared experts' marks against LSA marks using a gold standard	5	short answer	1,000	best result about .45	0.72		
LD01	compare with teacher - provided topic list	0-20	essay	31		0.59	no sig difference between 3 groups - 1 - control - no help 2 - human help provided; 3 - Apex help	
QKG01a	compare LSA with human assessment		moves in Microworld			0.57		
BB03	uses 2 similarity measures							
KKP03	used 20 good answers to each of 8 questions; correlation coefficient, MAD, correct vs false evaluations	2		192		0.47		
NVA03								
THD04	use Spearman's rho correlation to compare average human grade with LSA grade	8,2,7	essay	18			only 1 set was correlated statistically	
PGS05	Pearson correlation coefficient between humans' scores and Atenea's scores		short essays				0.5	not clear