The Open University

Seeing the Whole Picture: Comparing Computer Assisted Assessment Systems using LSA-based Systems as an Example

Debra Trusso Haley
Pete Thomas
Marian Petre
Anne De Roeck

15th June 2007

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

Seeing the Whole Picture: Comparing Computer Assisted Assessment Systems using LSA-based Systems as an Example

Debra Trusso Haley, Pete Thomas, Marian Petre, Anne De Roeck
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 presents a framework for evaluating computer assisted assessment (CAA) systems. It discusses why the framework can be useful for both producers and consumers of these automatic aids to assessing learners. The framework builds on previous work to analyse Latent Semantic Analysis- (LSA) based systems, a particular type of CAA, that produced a research taxonomy that could help LSA CAA developers to 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 a CAA 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 LSA-based CAA 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 CAA system.

1 Introduction

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

We are interested in providing tools, such as formative online tests, that improve the learning of programming and computing in general. 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 a Latent Semantic Analysis-based CAA (D. Haley, Thomas, De Roeck, & Petre, 2007) to mark short answers about html and other areas in computer science. 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.

Our work with LSA 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 with LSA, it is difficult to get a full picture of these systems. Some of the missing information includes type of training corpus and examples of questions being marked as well as the fundamental LSA options such as weighting function and number of dimensions in the reduced matrix.

Evaluation of the marking systems is a crucial topic because automatic marking systems will not be used if people do not have faith in their 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 these CAA systems.

The state of knowledge about CAA would be improved if researchers were able to share each others' experience in a meaningful way. It is difficult to compare these systems because no uniform procedure exists

for reporting CAA results. This paper attempts to fill that gap by proposing a framework for reporting on and evaluating automatic assessment tools.

1.1. Contribution

The major contribution of this paper is a two-part framework for comparing computer assisted assessment (CAA) systems. It is based on a research taxonomy (D. T. Haley et al., 2005) developed to compare Latent Semantic Analysis (LSA) based educational applications. The framework can be of value to both producers and consumers of CAA.

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 the system. If producers of CAAs use the framework, they can contribute to the improvement of CAA state-of-the-art by adding to a collection of comparable data.

Consumers are organisations, such as universities, that wish to use a CAA system. CAA consumers are, or should be, particularly interested in two areas. The most important area is the accuracy of the results. But what does accuracy mean and how do we measure it? We believe that a CAA system is *good enough* if its marks correlate to human markers as well as human markers correlate with each other. We discuss various ways to measure accuracy in previous work (D. Haley et al., 2007).

1.2. The Framework

The first part of the framework, which is for describing a CAA, can be visualised as the jigsaw puzzle in Figure 1. Figure 2 shows the second part of the framework - the evaluation of the CAA. We contend that all the pieces of this puzzle must be present if a reviewer wants to see the whole picture.

The important categories of information for specifying a CAA are the items assessed, the training data, and the algorithm-specific technical details. The general type of question (e.g., essay, 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 easier to mark 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 CAA. The central piece of Figure 1 shows LSA-specific options, but these could be changed if the CAA is based on a different method. The corpus used to train the CAA 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 CAAs. 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. Information about both these types of training data should be included.
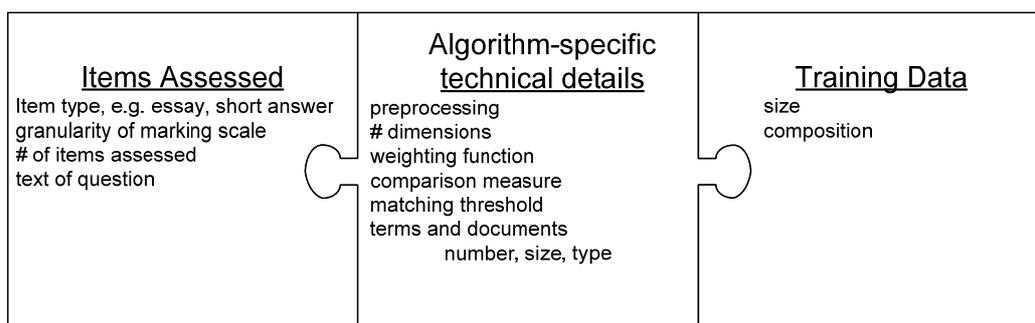
Figure 1. First part of framework: comparing Computer Assisted Assessment systems

Anyone interested in developing or using a CAA will be interested in its evaluation. The accuracy of the marks is of primary importance. A CAA 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, 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 Computer Assisted Assessment systems

1.3. Organization of the paper

Section 2 gives background on the development of the taxonomy from which the framework was devised. Section 3 shows how the framework can be used thro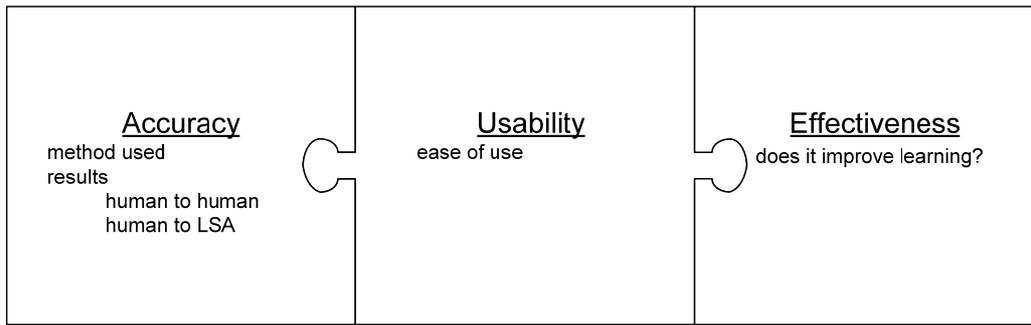ugh an example of our LSA-based CAA system (which was used to mark questions about html) and discusses the lessons learned from using the framework. Finally, Section 4 concludes by suggesting how the framework can be expanded from LSA-based CAAs to other types of assessment systems. (This paper does not offer an LSA tutorial. Readers desiring a basic introduction to LSA should consult the references section.)

2 Research taxonomy for LSA-based CAAs

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.

2.1. 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 the 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 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.

2.2. 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.
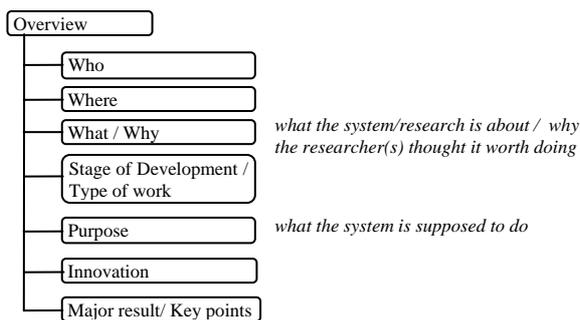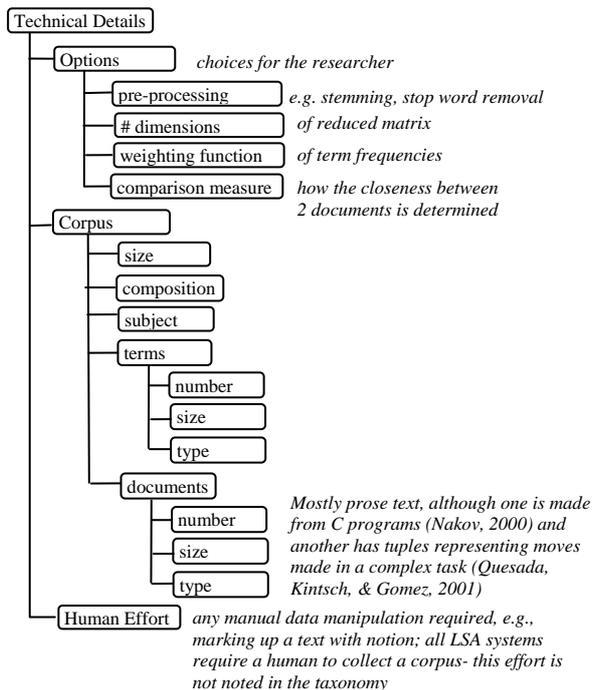
Figure 3. Category A: Overview

Figure 4. Category B: Technical Details

[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.

Figure 5. Category C: Evaluation

The figure shows a tree:

- Evaluation
  - accuracy — *these categories apply if the system assesses some kind of artefact; otherwise, the cells are shaded out*
    - method used
    - granularity of marks — *the finer the granularity the harder it is to match human markers*
    - item of interest — *e.g. essay, short answer*
    - # items assessed — *# students x # questions on exam*
    - results — *a successful LSA-based system should correlate to human markers as well as they correlate to each other*
      - human to LSA correlation
      - human to human correlation
  - effectiveness — *whether or not student learning is improved*
  - usability — *ease of use*

Appendix A presents the taxonomy. When looking at it, the reader should keep a few points in mind. First, 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. Second, the shaded cells indicate that the data item is not relevant for the article being categorised. Third, blank cells indicate that we were unable to locate the relevant information in the article. Fourth, 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).

3 Using the Framework for an LSA-based CAA

Table 2 is an example of how the framework could be used to compare different CAAs in tabular form. It starts with an overview and proceeds with the pieces in the puzzles of Figures 1 and 2. Obviously, it is very similar to the original taxonomy because it is based on it. The experience of creating and using it served to crystallize our thinking about the important elements of reporting on a CAA.

System Name	Reference	OverView						Items Assessed				Algorithm-specific Technical Details					Terms			Documents		
		Who / Where	What / Why	Stage of Development/ Type of work	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	Number	Size	Type	Number	Size	Type
EMMA	HTD 07	Thomas, Haley, De Roeck, Petre; The Open University	assess computer science short answers for summative assessment	research prototype	demonstrated the use of a research taxonomy for CAA, marked questions about html	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: <I>It is <B>very</I> </B> important to read this text carefully. — The desired appearance: It is *very* important to read this text carefully.)	stemming, stop words	90	log / entropy	cosine	none	12k	1 word	text	45k	1 paragraph	text
						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: Things to do: Pack suitcase, </BR> Book taxi. — The desired appearance: Things to do: Pack suitcase, Book taxi.)		500	log / entropy	cosine	none						

Reference	Training Data		Evaluation				
			Accuracy			Effectiveness	Usability
	Size	Composition	method used	Human to LSA	Human to Human	Does it improve learning?	How easy is it to use?
THD04	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	 54 32 11 1 1	not relevant - a research prototype	not relevant - a research prototype
THD04	1) 45k paragraphs 2) 80	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	 43 45 6 3 3	 61 28 9 1 1	not relevant - a research prototype	not relevant - a research prototype

Table 2. Filling in the framework

Our previous work (D. T. Haley et al., 2005) highlighted the insights revealed by the taxonomy. It listed the following major research themes:

- seminal literature describing the new technique named LSI, which was later renamed to LSA
- attempts to reproduce the results reported in the seminal literature, which for the most part failed to achieve the earlier results
- attempts to improve LSA by adding syntax information
- applications that analyse non-prose text.
- attempts to improve LSA by experimenting with corpus size and composition, weighting functions, similarity measures, number of dimensions in the reduced LSA matrix, and various pre-processing techniques

The major result of using the taxonomy is the conclusion 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.

4 Conclusions

We have confidence that our reporting and evaluation framework could support further research into LSA-based CAA tools by facilitating the sharing and comparison of results. 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, we believe that the same benefits that accrue to LSA researchers by using the framework can also extend to broader CAA 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 CAA producers and consumers. The ultimate goal is to improve computing education by improving assessment.

Acknowledgements

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 *Proc. 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 *Proc. 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. Psotka, 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 *Proc. 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., Rodriquez, 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 *Proc. 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 LSA Research Taxonomy

System Name	Reference	A1 Who	A2 Where	A3 What / Why	A4 Stage of Development/ Type of work	A5 Purpose	A6 Innovation	A7 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
	BDO95	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	Foltz, Britt, 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	representation generated by LSA is sufficiently similar to the readers' situation model to be able to characterize the quality of their essays
	FKL98	Foltz, 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 space	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, Render, Laham, Foltz, Kintsch,	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.nmsu.edu.	FLL99	Foltz, 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	A1 Who	A2 Where	A3 What / Why	A4 Stage of Development/ Type of work	A5 Purpose	A6 Innovation	A7 Major Result / Key points
Summary Street http://www.k-	KSS00	Kintsch, Steinhart, Stahl, LSA Research Group, Matthews, Lamb	U of Colorado, Platt Middle School	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 http://www.k-a-	Ste01	Steinhart	U of Colorado	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	
	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
	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,	U of Memphis, CHI Systems	evaluate corpora with different specificities for	LSA research for formative	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	provide feedback on topic, outline and coherence		LSA is a promising method to grade essays
Distribute d LSI	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.
	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
SELSA	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	SELSA has limited improvement over LSA
indexing not assessing	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, DeRoeck, Petre	The Open University	assess computer science essays	LSA research for summative assessment	assess essays	used a very small, very specifc 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, Rodriquez, Magnini	U de Madrid; Istituto per la Ricerca Scientifica e Technologica	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

	B1: Options				B2: Corpus									B3
	B1.1	B1.2	B1.3	B1.4	B2.1	B2.2	B2.3	B2.4: Terms			B2.5: Documents			
Refer-ence	pre-processing	# dimensions	weighting function	comparison measure	size	composition	Subject	B2.4.1 Number	B2.4.2 Size	B2.4.3 Type	B2.5.1 Number	B2.5.2 Size	B2.5.3 Type	Human Effort
DDF90	remove 439 stop words (from SMART)	100		cosine		MED	medical abstracts	5,823	words		1,033	average 50	title and abstract	
		100		cosine		CISI	information science abstracts	5,135	words		1,460	avg 45 words		
Dum91	remove 439 stop words (from SMART)	60, 100	log entropy			MED, CISI, CRAN, TIME, ADI	various (described in paper)	374 - 5831	words		82 - 1460		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		
		100		cosine		21 articles on the the heart	heart	2,781	words	prose text				
LD97		300	ln(1+freq)/entropy	cosine	4.6M	Grolier's Academic American Encyclopedia		60.7k	word	prose	30.4k	average 151	words	
LLR97	remove 439 stop words	94		cosine vector length		27 articles from Grolier's Academic Amer. Encyclopedia	heart anatomy	3034	word	prose	830	sentence	words	
	no stop words	1500				textbook	psychology	19,153	words	prose	4,904	paragraphs	words	
RSW98				cosine		27 articles fromGrolier's Academic Amer. Encyclopedia	heart anatomy							separated essays into technical and non technical
														created subsections of essays
WSR98		100		cosine	17,880	36 encyclopedia articles	heart	3,034	word	prose				
FLL99						a portion of the textbook	psycholinguistics							
							standardised test - opinion essays							
							standardised test - argument essays							
							diverse							
							heart							

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
							Meso-American history	46,951	1 word	prose text	530		prose text	
Ste01	correct spelling			cosine		general knowledge space	sources of energy		1 word	prose text				
				cosine		specialized texts	heart and lung							
				cosine		specialized texts	Meso-American history							
Lan02		300		cosine										
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"
WG00														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			religious texts	religion	20,433			196			
						C programs				C code				
NPM01	removed stop words and those occuring only once	15	6 different	dot product / cosine	974 K	Huckleberry Finn and Adventures of Sherlock Holmes		5534	words	prose	487	2 KB	prose	
FKM01		300		cosine		physics text book and other science text books	physics					paragraph	paragraph	prepare specialised corpora
OFK02		300		cosine		physics text book + related to curriculum script	physics	from 1,564 to 6,536	word	prose	from 416 to	paragraph	prose	sanitize corpora; write "expectations" for each
LD01					290K + size of course	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	3441	1 trial		
BB03							various							create a classification scheme for LSI vector
KKP03	removed stop words		log entropy	cosine	2.3M	used Auto tutor corpus	computer literacy	9,194	word	word - part of speech tags	5,596	paragraph	prose	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 paragraph	prose	none
PGS05			tf-idf			10 different corpora: student answers + text from popular computer magazines								

Reference	C1: Accuracy						C2	C3
	C1.1 Method used	C1.2 Granularity of marks	C1.3 Item of Interest	C1.4 Number items assessed	C1.5: results		Effectiveness	Usability
					C1.5.1 Human to LSA correlation	C1.5.2 Human to Human correlation		
DDF90	evaluate using recall and precision		queries	30				
			queries	35				
Dum91	evaluate using recall and precision							
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	5	short essay - 250 words	94	0.77	0.77		
	gold standard - a short text written by an expert			94	0.72			
	compare against human graders			273	0.64	0.65		
RSW98	compare with 1 or more target texts		short essay - 250 words	106				
WSR98	compared with 4 texts of increasing difficulty and specificity	5 point scale	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		
	holistic			188	0.8	0.83		
KSS00	compare with teacher - provided topic list	10	summary of essay				no sig difference	in classroom 1997-1999; students like immediate feed back
		10	summary of essay	50	0.64	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		52				
		10		52				
Lan02	holistic, Pearson product-moment correlation coefficient	5 or 10 points	essay	3,500	0.81	0.83		
WWG99	compare against pre-graded answers for completeness and compatibility	2: threshold of .55	short answers average length is 16 words	192	0.49	0.51		
Wie00	compared tuples in student answer with tuples in expected answer				.18, .24, and .4			
WG00								
WZ01	evaluate two texts using cosine							
Nak00 b	created correlation matrices							
NPM01	defined precision as ration of chunks from same text to num of chunks at a level							
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	0.68	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	0.59	
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	