

A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications

Debra Trusso Haley, Pete Thomas, Anne DeRoeck, Marian Petre

The Computing Research Centre

The Open University

Walton Hall, Milton Keynes MK7 6AA UK

[D.T.Haley, P.G.Thomas, A.DeRoeck, M. Petre] at open.ac.uk

Abstract

The paper presents a taxonomy that summarises and highlights the major research into Latent Semantic Analysis (LSA) based educational applications. The taxonomy identifies five main research themes and emphasises the point that even after more than 15 years of research, much is left to be discovered to bring the LSA theory to maturity. The paper provides a framework for LSA researchers to publish their results in a format that is comprehensive, relatively compact, and useful to other researchers.

1 Introduction

The major contribution of this paper is a taxonomy resulting from an in-depth, systematic review of the literature concerning latent semantic analysis (LSA) research in the domain of educational applications. The taxonomy presents the key points from a representative sample of the literature. Researchers and developers implementing LSA-based educational applications will benefit by studying the taxonomy because it brings to one place the techniques and evidence reported in the vast LSA literature.

We realized the need for a taxonomy while building an LSA-based assessment system for use in computer science courses. Although our original assessment results were encouraging, they were not good enough for the intended task of summative assessment (Thomas, Haley, et al. '04). We conducted a comprehensive, in-depth literature review to find techniques to improve our system. The taxonomy documents our findings and supports the insights gained by studying the literature.

There exists a great deal of literature on LSA. Some of it involves educational uses (Steinhart '01), some concentrates on LSA theory (Landauer & Dumais '97), and some of the newer articles¹ suggest uses of LSA that go beyond analysing prose

(Marcus, Sergeyev, et al. '04, Quesada, Kintsch, et al. '01).

The literature demonstrated that others were having difficulty matching the results reported by the original LSA researchers. We found a lot of ambiguity in various critical implementation details (e.g. weighting function used) as well as unreported details. We speculate that the conflicting or unavailable information explains at least some of the inability to match the success of the original researchers.

This paper is not an LSA tutorial. Readers desiring a basic introduction to LSA should consult the references section.

Section 2 explains the taxonomy, section 3 discusses insights gained by studying the taxonomy, and section 4 concludes with a suggestion for other LSA researchers.

Space limitation preclude presenting the taxonomy. See the Open University Technical Report 2005/09 at <http://computing-reports.open.ac.uk/> for the full, six page taxonomy.

2 About the taxonomy

2.1. Scope of the taxonomy

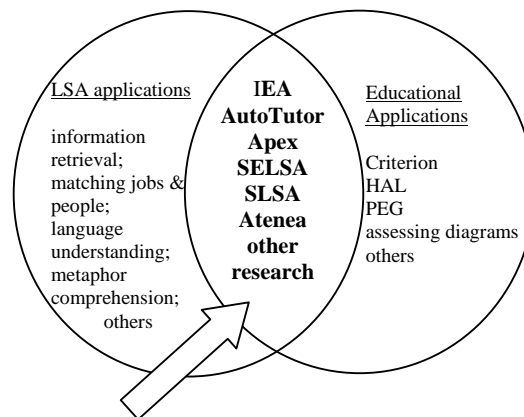


Figure 1. Scope of the Taxonomy – the intersection of LSA and educational applications

¹ To avoid confusion, we refer to papers in the literature as *articles*. *Paper* refers to this paper, which includes a taxonomy.

The taxonomy summarises and highlights 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. Therefore, LSA research into such areas as information retrieval (Nakov, Valchanova, et al. '03) and metaphor comprehension (Lemaire & Bianco '03) do not appear in the taxonomy. Similarly, the taxonomy ignores various non-LSA techniques that have been used to assess essays (Burgess, Livesay, et al. '98, Burstein, Chodorow, et al. '03) and diagrams (Anderson & McCartney '03, Thomas, Waugh, et al. in press).

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

2.2. Method for choosing articles

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 citee – 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, Deerwester, et al. '88), to one published in May 2005 (Perez, Gliozzo, et al. '05). 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).² Although not strictly in the scope of the intersection of LSA and educational applications, some of these articles appear in the

² 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.

taxonomy 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 present no new information pertinent to the taxonomy. This left us with 28 articles in the taxonomy.

2.3. 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 2, 3, and 4 show the headings and sub-headings. Most of the headings are self-explanatory; some clarifications are noted in the figures.

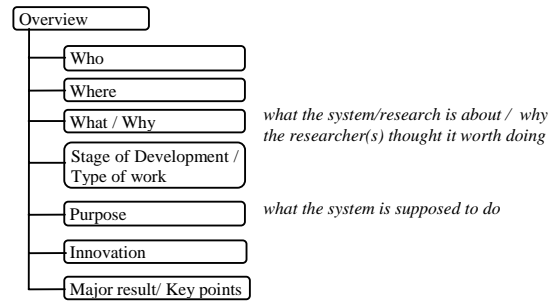


Figure 2. Category A: Overview

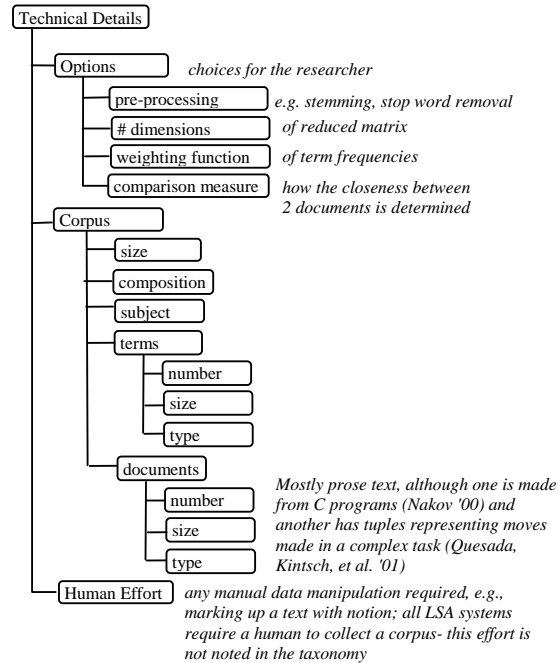


Figure 3. Category B: Technical Details

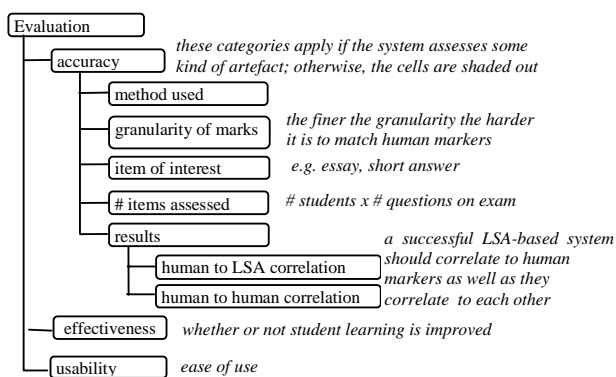


Figure 4. Category C: Evaluation

When looking at the taxonomy, 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.³ 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.⁴

Organising a huge amount of information in a small space is not easy. The taxonomy in the technical report (<http://computing-reports.open.ac.uk>) is based on an elegant solution in (Price, Baecker, et al. '93).

3 Discussion

This section discusses the insights revealed by the taxonomy. Sections 3.1 and 3.2 describe what can be found in the literature, and section 3.3 highlights some of the gaps in the literature.

3.1. Main research themes

A great deal of literature exists about LSA and about educational applications. Even the intersection of these two areas contains many articles. However, the taxonomy reveals five main research themes:

³ Please send any corrections to the first author, who will gladly update the taxonomy.

⁴ The *Reference* column contains a pointer to the references section at the end of this paper. Each reference contains a code at the end that corresponds to the entry in the *Reference* column. The entries are of the form *xxxnn* where *xxx* are the initials of up to three of the authors. If capitalised, they represent different authors; if the first is capitalised and the second two are lower case, the article has one author. *nn* is the 2-digit year of publication.

- 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 – exactly those items in Category B1 of the taxonomy

3.2. Diversity in the research

The taxonomy reveals a great deal of variety in the research. Researchers work in North America, Europe, and Asia on both deployed applications and continuing research. They use a wide variety of options for pre-processing techniques, number of dimensions in the reduced matrix, weighting functions, and composition and size of corpus. They use English, French, Spanish and Bulgarian corpora. The researchers report their evaluation methods with different specificity.

3.3. Gaps in the literature

The great variety of techniques used by researchers mentioned in the previous section leads to difficulty in comparing the results. Other researchers need to know all of the details to fully evaluate and compare reported results.

Much information is missing on page 2 of the taxonomy – *Category B: Technical Details*. These missing data concern the choices researchers must make when they implement their systems. Page 3 of the taxonomy, *Category C: Evaluation*, shows that some researchers have not evaluated the effectiveness or usability of their deployed systems.

The *Method used* subheading under *Accuracy* in *Category C* is a major area for gaps. Although many researchers report correlations between LSA and human graders, they usually do not mention whether they are using the Pearson, Spearman, or Kendall's tau correlation measure.

The existence of the blank cells in the taxonomy is troubling. They imply that researchers often neglect to report critical information, perhaps due to an oversight or page length restrictions. Nevertheless, the ability to reproduce results would be enhanced if more researchers provided more detailed data regarding their LSA implementations.

4 Conclusions

We hope that future LSA researchers will keep the taxonomy in mind when presenting their work. Using it will serve two main purposes. First, it will be easier to compare various research results. Second, it will ensure that all relevant details are provided in published articles, which will lead to improved understanding and the continued development and refinement of LSA.

The variability in the results documented in the taxonomy shows that LSA is still something of an art. More than 15 years after its invention, the research issues suggested by (Furnas, Deerwester, et al. '88) are still very much open.

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

- (Anderson & McCartney '03) M. Anderson & R. McCartney, Diagram processing: Computing with Diagrams. *Artificial Intelligence*, vol. 145, pp. 181-226, 2003 [AM03].
- (Bassu & Behrens '03) D. Bassu & C. Behrens. 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. pp, San Francisco, 2003 [BB03].
- (Berry, Dumais, et al. '95) M. W. Berry, S. T. Dumais & G. W. O'Brien, Using linear algebra for intelligent information retrieval. *SIAM Review* 37, vol. 4, pp. 573-595, 1995 [BDO95].
- (Burgess, Livesay, et al. '98) C. Burgess, K. Livesay & K. Lund, Explorations in context space: Words, sentences, discourse. *Discourse Processes*, vol. 25, pp. 211-257, 1998 [BLL98].
- (Burstein, Chodorow, et al. '03) J. Burstein, M. Chodorow & C. Leacock. 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*. pp, Acapulco, Mexico, 2003 [BCL03].
- (Deerwester, Dumais, et al. '90) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer & R. Harshman, Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990 [DDF90].
- (Dumais '91) S. T. Dumais, Improving the retrieval of information from external sources. *Behavioral Research Methods, Instruments & Computers*, vol. 23, pp. 229-236, 1991 [Dum91].
- (Foltz, Britt, et al. '96) P. W. Foltz, M. A. Britt & C. A. Perfetti. Reasoning from multiple texts: An automatic analysis of readers' situation models. In *18th Annual Cognitive Science Conference*. pp 110-115, NJ, 1996 [FBP96].
- (Foltz, Kintsch, et al. '98) P. W. Foltz, W. Kintsch & T. K. Landauer, The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Process*, vol. 25, pp. 285-307, 1998 [FKL98].
- (Foltz, Laham, et al. '99) P. W. Foltz, D. Laham & T. K. Landauer. Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99*. pp, 1999 [FLL99].
- (Franceschetti, Karnavat, et al. '01) D. R. Franceschetti, A. Karnavat, J. Marineau, G. L. McCallie, B. A. Olde, B. L. Terry & A. C. Graesser. Development of Physics Text Corpora for Latent Semantic Analysis. In *Proc. of the 23rd Annual conference of the Cognitive Science Society*. pp, 2001 [FKM01].
- (Furnas, Deerwester, et al. '88) G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter & K. E. Lochbaum. 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, 1988 [FDD88].
- (Kanejiya, Kumar, et al. '03) D. Kanejiya, A. Kumar & S. Prasad. Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. In *Building Educational Applications Using Natural Language Processing*, Proc. of the HLT-NAACL 2003 Workshop. pp 53-60, 2003 [KKP03].
- (Kintsch, Steinhart, et al. '00) E. Kintsch, D. Steinhart, G. Stahl, C. Matthews & R. Lamb, Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*. [Special Issue, J. Psotka, guest editor], vol. 8, pp. 87-109, 2000 [KSS00].
- (Landauer '02) T. K. Landauer. On the computational basis of learning and cognition: Arguments from LSA. In *The Psychology of Learning and Motivation*. edited by B. Ross, 41, pp 43-84, New York, 2002 [Lan02b].
- (Landauer & Dumais '97) T. K. Landauer & S. T. Dumais, A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, vol. 104, pp. 211-240, 1997 [LD97].
- (Landauer, Foltz, et al. '98) T. K. Landauer, P. W. Foltz & D. Laham, An introduction to Latent Semantic Analysis. *Discourse Processes*, vol. 25, pp. 259-284, 1998 [LFL98].
- (Landauer, Laham, et al. '97) T. K. Landauer, D. Laham, B. Rehder & M. E. Schreiner. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*. pp 412-417, 1997 [LLR97].
- (Lemaire & Bianco '03) B. Lemaire & M. Bianco. Contextual effects on metaphor comprehension: Experiment and simulation. In *Proceedings of the 5th Int'l Conference on Cognitive Modeling (ICCM'2003)*. pp, Bamberg, Germany, 2003 [LB03].

- (Lemaire & Dessus '01) B. Lemaire & P. Dessus, A system to assess the semantic content of student essays. *J. of Educational Computing Research*, vol. 24, pp. 305-320, 2001 [LD01].
- (Marcus, Sergeyev, et al. '04) A. Marcus, A. Sergeyev, V. Rajlich & J. I. Maletic. An Information Retrieval Approach to Concept Location in Source Code. In *Proceedings of the 11th IEEE Working Conference on Reverse Engineering*. pp 214-223, Delft, The Netherlands, 2004 [MSR04].
- (Nakov '00) P. Nakov. Latent Semantic Analysis of Textual Data. In *Proceedings of the Int'l Conference on Computer Systems and Technologies*. pp, Sofia, Bulgaria, 2000 [Nak00b].
- (Nakov, Popova, et al. '01) P. Nakov, A. Popova & P. Mateev. Weight functions impact on LSA performance. In *Proc. of the EuroConference Recent Advances in Natural Language Processing (RANLP'01)*. pp, Tzigrav Chark, Bulgaria, 2001 [NPM01].
- (Nakov, Valchanova, et al. '03) P. Nakov, E. Valchanova & G. Angelova. Towards Deeper Understanding of the LSA Performance. In *Proc. of Recent Advances in Natural Language Processing*. pp 311-318, Borovetz, Bulgaria, 2003 [NVA03].
- (Olde, Franceschetti, et al. '02) B. A. Olde, D. R. Franceschetti, A. Karnavat & A. C. Graesser. 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, 2002 [OFK02].
- (Perez, Gliozzo, et al. '05) D. Perez, A. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodriguez & B. Magnini. Automatic Assessment of Students' free-text Answers underpinned by the combination of a Bleu-inspired algorithm and LSA. In *Proceedings of the 18th Int'l FLAIRS Conference*. pp, Clearwater Beach, Florida, 2005 [PGS05].
- (Price, Baecker, et al. '93) B. A. Price, R. M. Baecker & I. S. Small, A Principled Taxonomy of Software Visualization. *Journal of Visual Languages and Computing*, vol. 4, pp. 211-266, 1993 [PBS93].
- (Quesada, Kintsch, et al. '01) J. Quesada, W. Kintsch & E. Gomez, 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*, vol. 43-84, pp. 117-131, 2001 [QKG01a].
- (Rehder, Schreiner, et al. '98) B. Rehder, M. E. Schreiner, M. B. W. Wolfe, D. Laham, T. K. Landauer & W. Kintsch, Using Latent Semantic Analysis to assess knowledge: some technical considerations. *Discourse Process*, vol. 25, pp. 337-354, 1998 [RSW98].
- (Steinhart '01) D. J. Steinhart, Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis. Unpublished PhD Thesis, Department of Psychology, University of Colorado, Boulder, 2001 [Ste01].
- (Thomas, Haley, et al. '04) P. Thomas, D. Haley, A. De Roeck & M. Petre. E-Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study. In *Proc. of the eLearning for Computational Linguistics and Computational Linguistics for eLearning Workshop at COLING 2004*. pp 38-44, Geneva, 2004 [THD04].
- (Thomas, Waugh, et al. in press) P. Thomas, K. Waugh & N. Smith. Experiments in the automatic marking of ER-Diagrams. In *Proc. of ITiCSE 05*. pp, Lisbon, Portugal, in press [TWS05].
- (Wiemer-Hastings '00) P. Wiemer-Hastings. Adding syntactic information to LSA. In *22nd Annual Conference of the Cognitive Science Society*. pp 989-993, 2000 [Wie00].
- (Wiemer-Hastings & Graesser '00) P. Wiemer-Hastings & A. C. Graesser, Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, vol. 8, pp. 149-169, 2000 [WG00].
- (Wiemer-Hastings, Wiemer-Hastings, et al. '99) P. Wiemer-Hastings, K. Wiemer-Hastings & A. C. Graesser. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *Artificial Intelligence in Education*. pp, Amsterdam, 1999 [WWG99].
- (Wiemer-Hastings & Zipitria '01) P. Wiemer-Hastings & I. Zipitria. Rules for Syntax, Vectors for Semantics. In *Proc. of the 23rd Cognitive Science Conference*. pp, 2001 [WZ01].
- (Wolfe, Schreiner, et al. '98) M. B. W. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch & T. K. Landauer, Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, vol. 25, pp. 309-336, 1998 [WSR98].