# The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis?

**Brent A. Olde (baolde@memphis.edu)**
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

**Donald R. Franceschetti (dfrncsch@memphis.edu )**
Department of Physics, University of Memphis, CAMPUS BOX 523390
Memphis, TN 38152 USA

**Ashish Karnavat (akarnavat@chiinc.com)**
CHI Systems, Inc., 716 N. Bethlehem Pike, Suite 300
Lower Gwynedd, PA 19002 USA

**Arthur C. Graesser (a-graesser@memphis.edu)**
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

**and the Tutoring Research Group**

**Abstract**

Student responses to conceptual physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Expert evaluations of student answers to questions were correlated with LSA metrics of the similarity between student responses and ideal answers. We compared the adequacy of several text corpora in LSA performance evaluation, including the inclusion of written incorrect reasoning and tangentially relevant historical information. The results revealed that there is no benefit in meticulously eliminating the wrong or irrelevant information that normally accompanies a textbook. Results are also reported on the impact of corpus size and the addition of information that is not topic relevant.

## Introduction

AutoTutor is an intelligent tutoring agent that interacts with a student using natural language dialogue (Graesser, Person, Harter, & TRG, in press; Graesser, VanLehn, Rose, Jordan, & Harter, 2001). The tutor's interactions are not limited to single-word answers or formulaic yes/no decision trees. AutoTutor attempts to tackle the problem of understanding lengthy discourse contributions of the student, which are often ungrammatical and vague. AutoTutor responds to the student with discourse moves that are pedagogically appropriate. It is this cooperative, constructive, one-on-one dialogue that is believed to produce learning gains (Graesser, Person, & Magliano, 1995). One major component in the comprehension mechanism is the knowledge representation provided by Latent Semantic Analysis (LSA). LSA is a statistical, corpus-based natural language understanding technique that computes similarity comparisons between a set of terms and texts (Kintsch, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998).

The present study focuses on the domain of conceptual physics. It should be noted that most modern physics texts (such as Hewitt, 1998) devote considerable space to the historical evolution of physical concepts, the cultural context of physics, and its social impact. Some authors also devote appreciable space to discussing discarded theories and chains of reasoning that lead to incorrect conclusions. Thus, a significant fraction of the text found in a physics text may exemplify incorrect thinking.

The Tutoring Research Group at the University of Memphis has been concerned with the best strategy for selecting a corpus of texts when constructing an LSA space. A naive approach would be to gather a number of physics texts, and combine them into one corpus. However, there are some important, unexplored issues that must be addressed about this approach. What should be done about the text that was written to illustrate incorrect reasoning? Does the inclusion of historical information or peripherally related information strengthen or dilute the accuracy with which physics concepts are represented in the LSA space? In short, how much special preparation of the corpus is needed, if it is to provide a reliable representation of the physics that students are expected to learn?

In this paper, we provide a brief overview of LSA and how it is used in our tutoring system. Then we discuss a study designed to address the matter of corpus selection by systematically testing the kind of texts

needed for a training corpus. We discuss the implications of these results for tutoring systems in general.

## Latent Semantic Analysis

LSA has recently been successfully used as a statistical representation of a large body of world knowledge (Kintsch, 1998; Landauer & Dumais, 1997). LSA provides the foundation for grading essays, even essays that are not well formed grammatically or semantically. LSA-based essay graders assign grades to essays as reliably as experts in composition (Foltz, Gilliam, & Kendall, 2000). LSA has been used to evaluate the quality of student contributions in interactive dialogs between college students and AutoTutor. AutoTutor is a tutoring system in the domain of computer literacy and most recently physics (Graesser et al., in press; Graesser et al., 2001). The LSA module evaluates the quality of student answers to questions almost as reliably as graduate research assistants (Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, Harter, Person, & TRG, 2000; P. Wiemer-Hastings, K. Wiemer-Hastings, Graesser, & TRG, 1999). Having established the utility of LSA in evaluating the quality of student essays and contributions in a tutoring systems on a variety of topics, we are presently interested in exploring what qualities a useful LSA space must have.

LSA is a mathematical technique in which the information contained in the co-occurrences of words in a body of text is compressed into a set of vectors in N-dimensional space. The input to LSA is a word co-occurrence matrix M, where the individual elements Mij is the number of times that the ith word occurs in the jth document. A document is an arbitrarily defined unit, but normally is a sentence, paragraph, or section in a text; for this project we used paragraphs as our document size. The rows and columns of the matrix are then subjected to mathematical transformations that take into account the frequency of the words used in each of the documents (Berry, Dumais, & O'Brien, 1995; Landauer et al., 1998). Using the mathematical technique of singular value decomposition, the matrix is then expressed as the product of three matrices, the second of which contains the singular values on the diagonal. Changing all but the largest N singular values to zero sets the dimensionality N of the vector space representing the text. The matrices are then re-multiplied to produce a matrix of the same dimensions of the original matrix.

By removing the lowest of the singular values we are seem to be eliminating spurious co-occurrences and capturing a more accurate representation of the meaning of the text (Landauer & Dumais, 1997). The reduced number of dimensions is sufficient for evaluating the conceptual relatedness between any two bags of words. A bag is an unordered set of one or more words. The match (i.e., similarity in meaning, conceptual relatedness) between two bags of words is computed as the geometric cosine (or dot product) between the two associated vectors, with values that normally range from 0 to 1. LSA cosine values successfully predict the coherence of successive sentences in text (Foltz, Kintsch, & Landauer, 1998), the similarity between student answers and ideal answers to questions (Graesser, P. Wiemer-Hastings, et al., 2000; Wiemer-Hastings et al., 1999), and the structural distance between nodes in conceptual graph structures (Graesser, Karnavat, Pomeroy, Wiemer-Hastings, & TRG, 2000).

At this point, researchers are continuing to explore the strengths and limitations of LSA in representing world knowledge. For example, it is widely accepted that LSA is not equipped to handle syntax, word ordering constraints, Boolean expressions, negation, or other precise analytic expressions.

## Overview of AutoTutor

In order to fully understand how we use LSA in AutoTutor, it is beneficial to understand the framework in which it is used. Therefore, we briefly provide a general overview of the AutoTutor architecture. A more thorough description is provided in previous publications (Graesser, Person et al., in press; Graesser et al., 1999; Wiemer-Hastings et al., 1998). AutoTutor's style of tutoring was modeled after actual human tutoring sessions (Graesser et al., 1995). The tutor starts out by asking a question or posing a problem that requires a paragraph-length answer. The tutor then works with the student to cover the essential points that the tutor deems necessary to adequately understand the answer to the question. When a question is answered, the process is repeated for a subsequent question. Since most human tutors are peers of the students, they are not what one would label as experts. Thus, they typically have a limited understanding of what the students are trying to convey, yet, they can typically determine whether a response is "in the ball park". Despite the lack of complete understanding, survey studies have shown a sizable advantage for face-to-face tutoring sessions over classroom situations (Cohen, Kulik, & Kulik, 1982).

The user interface for AutoTutor attempts to recreate this face-to-face environment. It consists of four windows: one for presenting the main question, a second for displaying animated or static graphics (simulating diagrams or drawings that a tutor might use to illustrate points), a third with an animated conversational agent, and a fourth for the student to type a reply. AutoTutor's animated agent has synthesized speech, a head, hands, and can be seen from the chest up. These features were designed to provide appropriate speech, facial reactions, and hand

gestures so the student gets both verbal and visual feedback in order to enhance and more appropriately mimic a one-on-one tutoring environment.

AutoTutor's knowledge of its tutoring domain resides in a curriculum script. This is a list of the questions or problems that the tutor is prepared to handle in a tutoring situation, along with good answers to the questions and problems (Putnam, 1987). A major portion of the script is the LSA space; it gets created from an assortment of texts collected from the domain of interest. This corpus is a set of general, non-specific information on the subject matter (e.g., a textbook on conceptual physics), plus specific information directly relevant to the curriculum script. This specific information is comprised of a relatively lengthy, complete, "ideal" answer. This complete answer is separated into a set of specific good answers which address one aspect of the ideal answer; these are sometimes called expectations or points. There are also a set of bad answers and how they would be corrected. Finally, for each expectation in the ideal answer, there are hints, prompts, and assertions that help the student construct an appropriate answer. There are a variety of other dialog moves and slots in the curriculum script that need not be addressed in the present study.

It is important to mention that the LSA corpora investigated in the present study included the general information from textbooks, but never included the question specific information. Thus, only the general physics information was trained in the LSA space. It could be argued that an LSA space should not have any trouble accounting for the content in the curriculum scripts (even if it was a small script) if the material included in the corpus was tailored specifically to the problems. Therefore, we are exploring how far we can go by exclusively focusing on the general content of physics, as manifested in a textbook on conceptual physics.

So how does AutoTutor use LSA during the tutorial interaction? Using the LSA derived cosine matches, AutoTutor evaluates the quality of the student's contributions within a conversation turn and across turns with respect to expected good answers and bad answers to the question. Based on values of these cosine matches, appropriate dialog moves are executed, such as feedback (positive, negative, neutral), pumps, prompts for specific words, hints, assertions, summaries, corrections, and follow-up questions. The smoothness of the mixed-initiative dialog in AutoTutor critically depends on the fidelity of the LSA space. This of course motivated us to test the performance of the LSA space on various tasks and measures.

## Methods

**Participants.** Participants were 120 students from The University of Memphis and Rhodes College; 80 of the students where non-physics majors and 40 were physics majors. Each participant answered 10 problems that were randomly selected from a set of 53 physics problems. Four physics experts answered all 53 questions and graded all answers on a standard 5-point grading scale (A, B, C, D, F). The interrater reliability of the experts was $r$ = .72. In the performance tests of LSA, we compared the expert ratings of the student answers to the LSA cosine scores. The LSA cosine scores are a match between the student answer and the ideal answer (i.e., answers created by the experts). The 4 experts had graduate degrees in physics (2 masters and 2 doctoral).

**Materials.** We have assembled five different physics corpora to test the effect of the content of the subject matter on the quality of the LSA solutions. The documents in the texts were classified into different rhetorical categories, such as exposition, example problems, historical material, incorrect reasoning, and so on. The fundamental research question is whether the inclusion of different texts and the resulting purity of content will have an impact on the tests of LSA performance.

All the corpora include text materials from the mechanics portion of Paul Hewitt's *Conceptual Physics* (1998). This text is widely used in conceptual physics courses at the college level. Our largest corpus, designated as "All", included chapters 2-9 of the Hewitt book plus six volumes of a comprehensive text aimed at students in technical or life science majors, two advanced texts in electromagnetism, and another two physics texts that were available electronically, a general text by Benjamin Crowell and more advances text by Frank Firk. A somewhat smaller corpus (designated as "Hewitt-Crowell (6)") was constructed from the former by deleting four of the texts; these texts were considered peripherally related to our conceptual physics domain because they were advanced texts mainly dealing with electromagnetism rather than mechanical physic. An even smaller corpus (designated "Hewitt-Crowell (2)") was created by further deleting chapters that did not cover mechanics. Next, we deleted any material from the remaining text that was identified by a physics professor as being primarily historical or involving misconceptions. This was our sanitizing procedure and resulted in the "Hewitt-Crowell (2-Sanitized)" corpus. Finally in the "Hewitt (Sanitized)" corpus, we included only those texts from Hewitt that had been sanitized. It should be noted that each of the successively refined or sanitized corpora was a proper subset of the preceding one. Table 1 summarizes the composition of the five corpora in addition to reporting

the number of paragraphs and the number of unique terms.

Table 1. List of five physics corpora via the chapters that comprise them. Columns with triangles signify sanitized corpora while squares signify unsanitized corpora.

| Texts | Hewitt Sanitized | Hewitt Crowell (2-Sanitized) | Hewitt Crowell (2) | Hewitt Crowell (6) | All |
|---|---|---|---|---|---|
| Linear Motion | ◢ | ◢ | ■ | ■ | ■ |
| Nonlinear Motion | ◢ | ◢ | ■ | ■ | ■ |
| Newton's Laws of Motion | ◢ | ◢ | ■ | ■ | ■ |
| Momentum | ◢ | ◢ | ■ | ■ | ■ |
| Energy | ◢ | ◢ | ■ | ■ | ■ |
| Rotational Motion | ◢ | ◢ | ■ | ■ | ■ |
| Gravity | ◢ | ◢ | ■ | ■ | ■ |
| Satellite Motion | ◢ | ◢ | ■ | ■ | ■ |
| Newtonian Physics | | ◢ | ■ | ■ | ■ |
| Conservation Laws | | ◢ | ■ | ■ | ■ |
| Modern revolution in physics | | | | ■ | ■ |
| Vibrations and Waves | | | | ■ | ■ |
| Electricity and Magnetism | | | | ■ | ■ |
| Optics | | | | ■ | ■ |
| Essential Physics | | | | | ■ |
| Electromagnetic | | | | | ■ |
| Field Theory | | | | | ■ |
| Electrostatics and Circuits | | | | | ■ |
| Number of Paragraphs | 416 | 698 | 2051 | 3445 | 3778 |
| Number of Terms | 1564 | 2183 | 4169 | 6139 | 6536 |

**Measures.** The performance measure was computed on the set of answers to the 53 questions. Since there were 53 questions and approximately 20 answers per question, there was a set of approximately 1000 answers. Each answer was rated by the 4 experts on a 5 point scale (1 = F and 5 = A); the final grade for the answer was the mean grade of the 4 experts. We refer to this score as the grade of the answer. Also associated with each answer was an LSA coverage score, this score compared each student answer to the set of expectations in the experts' answers to the question. More specifically, each expert answers was segregated into a set of expectations, with each expectation being one sentence. An expectation was scored as "covered" if the LSA score between any sentence in the student answer and the expectation under consideration had an LSA cosine score that was greater than or equal to some threshold *T*. The extent to which student answer *S* matched expert answer *A* was computed as the proportion of expectations in *A* that had LSA matches that met the threshold (see Graesser et al., 2000). There were 4 of these scores, one for each of the 4 experts. The maximum value of these scores was designated as the LSA coverage score for student answer *S*. Moreover, we varied the thresholds in these computations from .3 to .9 in increments of .1 (see Figure 1). The correlation between the grades of the answers and the LSA coverage scores was the critical performance measure for the LSA space. The higher the correlation, the better the performance of the LSA space.

## Results and Discussion

We tested 5 different physics corpora, each having a slightly different level of specificity in the domain of conceptually based mechanical physics. Because the size of the corpus could affect the dimensionality and threshold, we tested the performance of all 5 levels of corpus size on 5 different dimensionalities (100, 200, 300, 400, and 500), and 7 critical threshold values, from 0.3 to 0.9 in 0.1 increments. For each combination of these factors, we computed the correlation between the grades and the LSA coverage scores.

Figure 1 plots performance for each level of corpus size by threshold at 300 dimensions. We used 300 dimensions for two reasons. First, the sanitized Hewitt corpus was so small that nothing higher than a 300 dimensional representation could be obtained. Second, the performance did not improve after 300 dimensions on any of the corpora. As Figure 1 shows, the LSA performance was practically identical for all corpus sizes except the smallest. Thus, it was not necessary to eliminate historical material, explanations of discarded theories, or useful demonstrations of incorrect chains of reasoning. There was no payoff in sanitizing the corpus.

The size of the corpus had a modest impact on the correlations, except for the extremely small corpus. Clearly the amount of text and the performance of LSA is not a linear relation. A relatively small amount of relevant material can produce acceptable performance with LSA.

According to the results in Figure 1, it appears that a threshold of approximately .8 provides a reasonable fit to the data. Thus, a sentence-like expectation is regarded as covered if there is a sentence in the student answer that has an LSA match score of .8 or higher.

In summary, we have developed a number of alternative physics text corpora for use in the evaluation of student answers to physics questions. Comparisons of the expert grades of the student answers and the computed LSA coverage scores suggest that the inclusion of material that is historical in nature or that exemplifies incorrect notions of physics does not hamper the performance the LSA space. It was also

surprising that the space performed as well as it did considering that there was no problem-specific information in the set of texts used for training the LSA space. Furthermore, a relatively small space in the restricted domain of physics contains enough information to mine an appropriate co-occurrence matrix and produce a properly functioning LSA space. Our current plan is to follow up this experiment by investigating how much performance is improved by adding the specific curriculum script information.

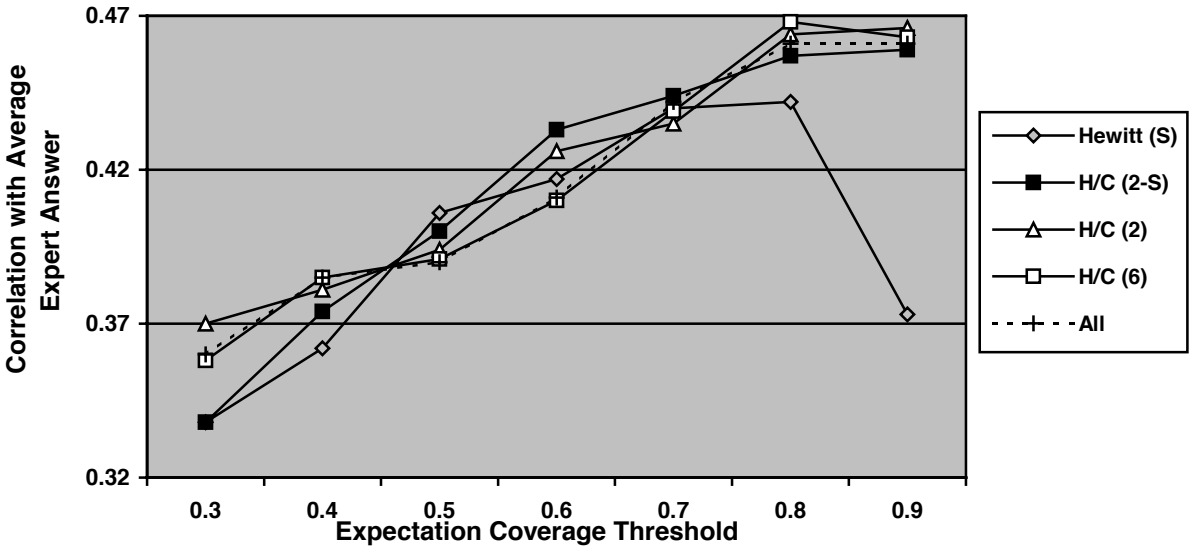


Figure 1: Correlation between the average expert grade and the student's LSA coverage score as a function of threshold and corpus of texts.

## Acknowledgments

This research was directly supported by the National Science Foundation (REC 0106965) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

## References

Albacete, P. L., & VanLehn, K. A. (2000), Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 25-30). Mahwah, NJ: Lawrence Erlbaum.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995), Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573-595.

Cohen, P. A., Kulik, J. A, and Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

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

Foltz, P. W., Kintsch, W., & Landauer, T. K, (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes 25*, 285-307.

Graesser, A. C., Karnavat, A., Pomeroy, A., Wiemer-Hastings, K., & TRG (2000), Latent semantic analysis captures causal, goal-oriented, and taxonomic structures. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 184-189) Mahwah, NJ: Erlbaum.

Graesser, A.C., Person, N., Harter, D., & TRG (in press). Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 359.1-28.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., & Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*, 128-148.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the TRG (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 35-51.

Hewitt, P. G. (1998) *Conceptual physics* (Ed. 8). Reading, MA: Addison Wesley Longman.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

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*, 211-240.

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

Ploetzner, R., & VanLehn, K. (1997). *Cognition & Instruction, 15*, 169-205.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*:13-48.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review or research-based instructional strategies, *American Journal of Physics, 59*, 891-897.

Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 334-343). Berlin, Germany: Springer-Verlag.

Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A. & TRG (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 535-542). Amsterdam: IOS Press.