AutoTutor's Coverage of Expectations during Tutorial Dialogue

Art Graesser, Andrew Olney, Matthew Ventura, and G. Tanner Jackson

Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152
a-graesser@memphis.edu, aolney@memphis.edu, mventura@memphis.edu, gtjacksn@memphis.edu

Abstract

AutoTutor is a learning environment with an animated agent that tutors students by holding a conversation in natural language. AutoTutor presents challenging questions and then engages in mixed initiative dialogue that guides the student in building an answer. AutoTutor uses latent semantic analysis (LSA) as a major component that statistically represents world knowledge and tracks whether particular expectations and misconceptions are expressed by the learner. This paper describes AutoTutor, reports some analyses on the adequacy of the LSA component, and proposes some improvements in computing the coverage of particular expectations and misconceptions.

Tutorial Dialogue with AutoTutor

AutoTutor is a computer tutor that holds conversations with learners in natural language (Graesser et al. 2004; Graesser, Person, & Harter 2001; Graesser, VanLehn, Rose, Jordan, & Harter 2001). AutoTutor simulates the discourse patterns of human tutors and is augmented with a number of ideal tutoring strategies. The tutor presents a series of challenging questions from a curriculum script and engages in a collaborative mixed initiative dialog while constructing answers. AutoTutor was designed to be a good conversational partner that comprehends, speaks, points, and displays emotions, all in a coordinated fashion. AutoTutor "speaks" by utilizing a speech engine developed at Microsoft (www.microsoft.com/products/msagent) or SpeechWorks (www.speechworks.com). For some topics and versions of AutoTutor, there are graphical displays, animations of causal mechanisms, or interactive simulation environments, with AutoTutor talking about and pointing to various components (Graesser, Chipman, Haynes, & Olney, in press). The initial versions of AutoTutor were on the topic of computer literacy. A later version, called *Why/AutoTutor*, helps college students learn Newtonian physics (Graesser, VanLehn et al. 2001) by asking them why-questions on difficult problems.

There are a number of distinctive features of AutoTutor that set it apart from most other intelligent tutoring systems. One strength of AutoTutor is that the tutoring domain can be changed quickly with lesson authoring tools, without the need to rebuild any of the conversational or pedagogical components of the system. A second strength is that the dialogue quality is sufficiently robust that bystander evaluators cannot tell the difference between a speech act generated by AutoTutor and a speech act generated by a human tutor; such bystander Turing tests have been reported in several experiments by Person and Graesser (2002). A third strength is that AutoTutor shows impressive learning gains of nearly a letter grade improvement compared with pretest measures and suitable comparison conditions (Graesser et al. 2003; Graesser, Lu et al. 2004; VanLehn, Graesser et al. 2004). A fourth strength is that AutoTutor uses a hybrid of structured knowledge representations and a statistical representation of world knowledge called Latent Semantic Analysis (LSA, Landauer, Folts, & Laham 1998; Graesser et al. 2000). It is this LSA component that will be examined in the present article.

Specific Goals and Scope of AutoTutor

AutoTutor presents a series of difficult questions to be answered, or problems to be solved. An answer to a typical question is 3-7 sentences, or approximately a paragraph of information. When students are asked questions that require paragraph-length answers and deep reasoning, initial answers to these questions are typically only 1 or 2 sentences in length. AutoTutor helps the student fill in missing pieces of the answer by managing a mixed initiative dialogue with different categories of dialogue acts. AutoTutor provides *feedback* to the student on what the student types in (positive, neutral, or negative feedback), *pumps* the student for more information ("What else?"), *prompts* the student to fill in missing words, gives the student *hints*, fills in missing information with *assertions*, identifies and *corrects* erroneous ideas and misconceptions, *answers* the student's questions, and *summarizes* answers. These dialogue acts of feedback, pumps, prompts, hints, assertions, corrections, answers, and summaries eventually lead to a full correct answer.

The tutorial dialog patterns of AutoTutor were motivated by research in discourse processing, cognitive science, and intelligent tutoring systems. Constructivist theories of learning emphasize the importance of learners actively constructing explanations (Aleven & Koedinger 2002; Chi, de Leeuw, Chiu, & LaVancher 1994; McNamara 2004). Researchers have developed intelligent tutoring systems that adaptively respond to the learner's knowledge and help construct explanations (Anderson Corbett, Koedinger, & Pelletier 1995; VanLehn et al. 2003). Empirical research in discourse processing has documented the

collaborative constructive activities that frequently occur during human tutoring (Chi, Siler, Jeong, Yamauchi, & Hausmann 2001; Fox 1993; Graesser, Person, & Magliano 1995).

Surprisingly, the dialog moves of most human tutors are not particularly sophisticated from the standpoint of today's pedagogical theories and intelligent tutoring systems. Human tutors rarely implement *bona fide* Socratic tutoring strategies, modeling-scaffolding-fading, reciprocal training, building on prerequisites, and other sophisticated pedagogical techniques. Instead, human tutors normally coach the student by filling in missing pieces of information in an expected answer and by fixing bugs and misconceptions expressed by the student. We refer to this tutoring mechanism as *Expectation and Misconception Tailored Dialog*. AutoTutor was designed to simulate the dialog moves of human tutors who coach students in constructing explanations and answers to open-ended questions.

A Concrete Example of AutoTutor

AutoTutor's curriculum script consists of a set of questions (or problems) and answers that require deep reasoning. Listed below are the important forms of content affiliated with each main question.

Main Question. If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Explain why.

Ideal Answer. The force of impact on each of the colliding bodies is due to interaction between them. The forces experienced by these bodies are thus an action/reaction pair. Thus, in terms of Newton's third law of motion, these forces will be equal in magnitude and opposite in direction. The magnitude of the acceleration produced by a force on different objects is inversely proportional to their masses. Hence, the magnitude of the car's acceleration due to the force of impact will be much larger than that of the more massive truck. A larger magnitude of acceleration implies a larger rate of change of velocity, which may be interpreted as greater change in motion. Therefore, the car undergoes greater change in its motion.

Expectations.
(E1) The magnitudes of the forces exerted by the two objects on each other are equal.
(E2) If one object exerts a force on a second object, then the second object exerts a force on the first object in the opposite direction.
(E3) The same force will produce a larger acceleration in a less massive object than a more massive object.

Misconceptions.
(M1) A lighter/smaller object exerts no force on a heavier/larger object.
(M2) A lighter/smaller object exerts less force on other objects than a heavier/larger object.
(M3) The force acting on a body is dependent on the mass of the body.
(M4) Heavier objects accelerate faster for the same force than lighter objects.
(M5) Action and reaction forces do not have the same magnitude.

Functionally Equivalent Concepts.
car, vehicle, object
truck, vehicle, object

The learner might possibly articulate the ideal answer to this question, but such a complete answer rarely if ever occurs. Natural language is much too imprecise, fragmentary, vague, ungrammatical, and elliptical to anticipate such semantically well-formed and complete responses. LSA is used to evaluate, probabilistically, the extent to which the information within the student turns (i.e., an individual turn, a combination of turns, or collective sequence of turns) matches the ideal answer. AutoTutor requires that the learner articulate each of the expectations before it considers the question answered. The system periodically identifies a missing expectation during the course of the dialogue and posts the goal of covering the expectation. When expectation E is missed and therefore posted, AutoTutor attempts to get the student to articulate it by generating hints and prompts that encourage the learners to fill in missing content words and propositions. Expectation E is considered covered if the content of the learners' turns meet or exceed a threshold T in its LSA cosine value. More specifically, E is scored as covered if the cosine match between E and the student input I is high enough: cosine (E, I) $\geq$ T. The threshold has varied between .40 and .75 in previous versions and evaluations of AutoTutor.

Each expectation E has a family of prompts and hints that potentially may be recruited in AutoTutor's dialogue acts in order to get the student to fill in every content word and proposition in E. A particular prompt or hint from the family is selected to maximize an increase in the LSA cosine *match score* when the prompt or hint is expressed by the learner. Students may express a misconception during the dialogue. This happens when the student input I matches a misconception M with a sufficiently high match score. AutoTutor corrects the misconception and goes on.

AutoTutor systematically manages the dialogue when it attempts to get the learner to articulate an expectation E that gets posted. AutoTutor stays on topic by completing the sub-dialog that covers expectation E before starting a sub-dialog on another expectation. Learners often leave out a content word, phrase, or entire clause within E. As

already mentioned, specific prompts and hints are selected that maximize the learner's filling in this content and boosting the match score above threshold. Suppose, for example, that expectation E1 needs to be articulated in the answer. The following family of candidate prompts is available for selection by AutoTutor to encourage the student to articulate particular content words in expectation E1.

(a) The magnitudes of the forces exerted by two objects on each other are ____.
(b) The magnitudes of forces are equal for the two _____.
(c) The two vehicles exert on each other an equal magnitude of _____.
(d) The force of the two vehicles on each other are equal in _____.

If the student fails to articulate one of the four content words (*equal, objects, force, magnitude*), then AutoTutor selects the corresponding prompt (a, b, c, and d, respectively).

LSA plays a critical role in AutoTutor in several respects. LSA serves as a conceptual pattern matcher as it constantly is comparing learner input to expectations and misconceptions. It provides a quantitative metric for evaluating the extent to which any two bags of words meet or exceed a threshold criterion. LSA is a statistical metric for performing pattern recognition, pattern matching, and pattern completion operations. The fidelity of the LSA component needs to be carefully evaluated in light of its central role in AutoTutor.

LSA-based Metrics that Evaluate the Coverage of Expectations and Misconceptions

This section reports some of the analyses that we have conducted in our evaluations of the extent to which LSA accurately assesses coverage of expectations and misconceptions. We have considered different units of analysis, spans of text, and algorithms in these assessments. These assessments of LSA have been conducted on the topics of computer science (Graesser, P.Wiemer-Hastings, K. Wiemer-Hastings, Person, & Harter 2000; P. Wiemer-Hastings, K. Wiemer-Hastings, & Graesser 1999) and Newtonian physics (Olde, Karnavat, Franceschetti, & Graesser 2002).

An LSA space requires a corpus of training texts, so a few words should be said about the corpora we used for AutoTutor. The versions of AutoTutor for computer literacy covered the topics of hardware, operating systems, and the Internet. The corpus consisted of a textbook on computer literacy, the curriculum script, and 30 articles, namely 10 for each of the three topics. The version of AutoTutor for physics consisted of the curriculum script, a textbook written by Hewitt (1998), and 10 articles on Newtonian physics. We used 300 dimensions in the LSA space for all of these tutors. It should be noted that the performance of LSA did not improve much when we sanitized the corpus by including only relevant documents and content (Olde et al. 2002), but we would expect performance to noticeably improve with a corpus that is 3-4 times the current size.

Evaluating Answer Essays

One of the performance assessments of AutoTutor is an essay test that is administered before and after AutoTutor training. The essays consist of students answering qualitative physics questions on their own, without any assistance. The major research question is how well LSA fares in grading these essays. Automated essay graders have adopted LSA metrics (Foltz, Gilliam, & Kendall 2000), so we pursued a similar assessment for AutoTutor. The standard approach is to compare LSA-based grades of essays with essay quality ratings of subject matter experts.

We have we asked experts in physics or computer literacy to make judgments about the overall quality of the student essays. An expert's *quality rating* was operationally defined as the proportion of expectations in an essay that judges believed were covered (using criteria that vary from stringent to lenient). Similarly, LSA was used to compute the proportion of expectations covered, using varying thresholds of match scores on whether information in the student essay covered each expectation. Correlations between the LSA scores and the judges' quality ratings (i.e., the mean rating across judges) were approximately .50 for both conceptual physics (Olde et al. 2002) and computer literacy (Graesser et al. 2000; P. Wiemer-Hastings et al. 1999). Correlations have generally increased as the length of the text increases, yielding correlations of .73 or higher in other laboratories (Foltz et al. 2000). We believe that our LSA-based assessments would exceed the .50 correlation if the answers had more words and the corpus was much larger. It is informative to note that the correlations between a pair of experts in our studies was approximately .65, so LSA agrees with experts approximately as good as experts agree with each other.

User Modeling During AutoTutor Training

As students contribute information, turn by turn, their content is compared with the expectations and misconceptions in the curriculum script. How well does LSA perform this user modeling? We have performed some analyses on the LSA cosine scores in AutoTutor's log files in order to answer this question.

In one analysis of conceptual physics, pretest scores on a multiple choice test served as the gold standard of the students' pre-experimental knowledge of physics before they started the AutoTutor training. The multiple choice test was similar to a frequently used test developed by

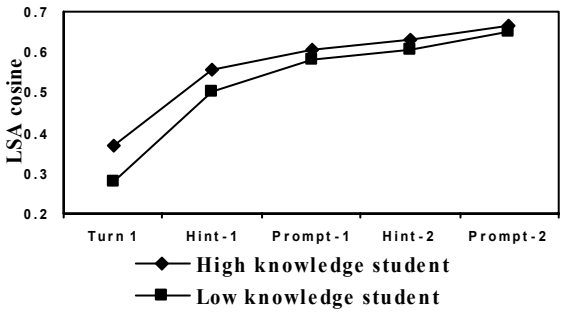
Hestenes, Wells, and Swackhamer (1992), called the Force Concept Inventory. If AutoTutor is performing effective user modeling, then the dialogue moves selected by AutoTutor should be systematically affected by the students' prior knowledge of physics. This indeed was the case when we analyzed the dialogue moves (Jackson, Mathews, Lin, & Graesser 2004).

Consider first the short feedback that AutoTutor gives to the student after most of the student's turns. The students' physics knowledge has a significant positive correlation with proportion of short feedbacks that were positive and a negative correlation with negative feedback. Next consider the corrections that AutoTutor made when identifying student errors and misconceptions. There was a negative correlation between student knowledge and the number of misconceptions identified by AutoTutor. Consider finally the three dialogue move categories that attempt to cover the content of the expectations in the curriculum script: Hints, prompts, and assertions. There is a continuum from the student supplying information to the tutor supplying information as we move from hints, to prompts, to assertions. The correlations with student knowledge reflected this continuum perfectly, with correlations being positive for hints and negative for assertions. For students with more knowledge of physics, all AutoTutor needs to do is pump and hint, thereby nudging the student to articulate the expectations. For students with less knowledge of physics, AutoTutor needs to generate prompts for specific words or to assert the correct information, thereby extracting knowledge piecemeal or telling the student the correct information. These results support the claim that AutoTutor performs user modeling with some modicum of accuracy; the system adaptively responds to the learner's level of knowledge.

Coverage Characteristics Curves. A different approach to assessing user modeling is by analyzing the evolution of answers to the main questions. As the student contributes information, turn, by turn, the LSA coverage scores should increase. *Coverage characteristics curves* (CC-curves) were prepared in order to assess whether this is the case. A CC-curve is the LSA score for an expectation E plotted as a function of conversational turns or as a function of particular states in the evolution of the answer. For example, Graesser et al. (2000) reported CC-curves for computer literacy by plotting mean cumulative LSA cosine scores for each expectation as a function of conversational turns. The scores were cumulative in the sense that the student's content in turn N+1 includes all of the content articulated by the student in turns 1 through N. The mean LSA scores were .27, .39, .45, .62, .66, and .76 for turns 1, 2, 3, 4, 5, and 6, respectively. These data support the claim that LSA adequately tracks the evolving evolution of a correct answer over conversational turns.

We prepared some CC-curves for one of our recent experiments on physics. We identified the total set of dialogue sequences in which the student had trouble articulating a particular expectation E. The LSA coverage score for E was recorded at 5 points: (a) after the question was first asked and AutoTutor pumped the student for information for one turn, (b) a first hint was given, (c) a first prompt was given (always after a first hint), (d) a second hint was given, and (e) a second prompt was given (always after the second hint). Figure 1 plots the mean LSA scores as a function of these five states in the dialogue history for an expectation E. As can be seen in Figure 1, there is a monotonic increase in LSA values over turns. The values are also higher for students with high than low pre-experimental knowledge about physics. Therefore, AutoTutor's LSA component provides a valid metric for tracking the coverage of an expectation over the tutorial dialogue.

Figure 1. Relationship between LSA cosine and number of student contributions in AutoTutor's physics tutor.



Matching Sentential Expectations to Learners' Sentential Contributions. An expert physicist rated the degree to which particular speech acts expressed during AutoTutor training matched particular expectations. These judgments were made on a sample of 25 physics expectations (sentence-length units, such as E1 through E5) and 5 randomly sampled student turns per expectation, yielding a total of 125 pairs of expressions. The learner turns were always responses to the first hint for that expectation. The question is how well the expert ratings correlate with LSA coverage scores.

The correlation between an expert judge's rating and the LSA cosine was modest, only $r = .29$. We scaled the 125 items on two other scales, to see how they would compare with LSA. First, we computed overlap scores between the words in the two sentential units (minus *a, an, the*). If an expectation has A content words, a student speech act has B content words, and there are C common words between the two sentential units, then the overlap score is computed as [2C/(A+B)]. The correlation between expert ratings and word overlap scores was $r = .39$. So a simple word overlap metric does a somewhat better job than LSA per se when sentential units are compared. Second, we scaled whether the serial order of common content words was similar between the two sentential units by computing Kendall's Tau scores. This word-order similarity metric had a .25

correlation with the expert ratings. We performed a multiple regression analysis that assessed whether the expert ratings could be predicted by LSA, word overlap, and Kendall's Tau together. The three predictors accounted for a significant r = .42. These results support the conclusion that analyses of sentences would benefit from a hybrid computational algorithm that considers both LSA and alternative algorithms. Alternative algorithms would consider (a) associates of the content words computed as nearest neighbors in the LSA space (Kintsch 2001), (b) word combinations in the actual corpus (Ventura et al. 2004), and (c) word sequences. A hybrid model between LSA and symbolic models of syntax and meaning would be a sensible research direction.

What Expectations are LSA-worthy?

It is conceivable that some expectations are more amenable to LSA computations than others. For example, expectations that have high frequency words should have poorer performance because there would not be enough distinctive content. Ideally, AutoTutor would have a principled way of determining whether or not an expectation is *LSA-worthy*. An expectation is defined as LSA-worthy if there is a high correlation between LSA and human experts in its being covered in essays or the training history. If AutoTutor could predict the LSA worthiness of an expectation in a principled way, then that could guide whether it trusted the LSA metrics. LSA metrics would be used for expectations that are LSA-worthy, but other algorithms would be used for expectations that are not LSA-worthy.

We computed LSA-worthiness scores for approximately 40 expectations associated with the physics problems that were used in the pretest and posttest essays. The LSA worthiness score for expectation E is simply a correlation between (a) the LSA coverage scores for E on a sample of essays and (b) the mean expectation coverage rating among 5 expert judges. These judges score whether any given expectation E was present in a particular essay in the sample. The correlation score between (a) and (b) should approach 1.0 to the extent an expectation is LSA-worthy. Some examples of these scores are listed for 4 of the expectations.

After the release, the only force on the balls is the force of the moon's gravity (r = .71)
A larger object will experience a smaller acceleration for the same force (r = .12)
Force equals mass times acceleration (r = .67)
The boxes are in free fall (r = .21)

The first and third of these expectations would be considered LSA-worthy, but not the second and fourth. It is fortunate that the third expectation is LSA-worthy because it captures Newton's second law.

We performed some analyses that assessed whether there were linguistic or conceptual properties of the expectations that would correlate with the LSA-worthiness scores. If there are, we would have a principled way for AutoTutor to gauge whether the LSA assessments can be trusted. We have examined dozens of linguistic and conceptual properties of the expectations, but only two of them significantly correlated with LSA-worthiness: Number of infrequent words (r = .23) and negations (r = -.29). Among the other properties that had near-zero correlations were number of words, content words, glossary terms, relative terms (e.g., *small, fast*), quantifiers (e.g., *some, all, one, two*), deictic expressions (e.g., *this, that*), and vector length. These modest correlations suggest it may be difficult to predict a priori which of the expectations will end up being LSA-worthy.

Conclusions about LSA

We are convinced that LSA has been moderately successful as a foundational representational system for AutoTutor. It was capable of grading essays in physics and computer literacy (r = .5), almost as well as experts. It could significantly perform user modeling and track the coverage of expectations during the evolution of collaborative dialogue. Sentence unit matches with LSA are limited (r = .3), however, so it is important to consider explicit word overlap and ordering of words when sentential units are compared. We know that some expectations can be reliably computed with LSA, but others cannot be and will require more hybrid architectures for conceptual pattern matching. We will explore some theoretically principled hybrid models in our quest for an adequate sentential pattern matcher.

Acknowledgements

The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of approximately 35 researchers from psychology, computer science, physics, and education (visit http://www.autotutor.org). The research on AutoTutor was supported by the National Science Foundation and the Department of Defense Multidisciplinary University Research Initiative administered by the Office of Naval Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DoD, ONR, or NSF.

References

Aleven, V., & Koedinger, K. R. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science* 26:147-179.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* 4:167-207.

Chi, M.T.H., de Leeuw, N., Chiu, M. & LaVancher, C. 1994. Eliciting self-explanation improves understanding. *Cognitive Science* 18:439-477.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. 2001. Learning from human tutoring. *Cognitive Science* 25:471-533.

Foltz, P.W., Gilliam, S., & Kendall, S. 2000. Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments* 8:111-127.

Fox, B. 1993. *The human tutorial dialog project*. Hillsdale, NJ: Erlbaum.

Graesser, A.C., Chipman, P., Haynes, B.C., & Olney, A. 2004. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. Submitted to *IEEE Transactions on Learning*.

Graesser, A.C., Hu, X., Person, P., Jackson, T., and Toth, J 2004. Modules and information retrieval facilities of the Human Use Regulatory Affairs Advisor (HURAA). *International Journal on eLearning*. Forthcoming.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*. Forthcoming.

Graesser, A.C., Person, N., Harter, D., & the TRG 2001. Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education* 12:257-279.

Graesser, A. C., Person, N. K., & Magliano, J. P. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology* 9:1-28.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine* 22(3):39-51.

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.

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* 8:129-148.

Hestenes, D., Wells, M., & Swackhamer, G. 1992. Force Concept Inventory. *The Physics Teacher* 30:141-158.

Hewitt, P. G. 1987. *Conceptual Physics*. Menlo Park, CA: Addison-Wesley.

Hume, G. D., Michael, J.A., Rovick, A., & Evens, M. W. 1996. Hinting as a tactic in one-on-one tutoring. *The Journal of the Learning Sciences* 5:23-47.

Jackson, G.T., Mathews, E.C., Lin, D., Graesser, A.C. 2003. Modeling student performance to enhance the pedagogy of AutoTutor. In P. Brusilovsky, A. Corbett, and F. de Rosis (Eds.), *Lecture Notes in Artificial Intelligence: 2702,* 368-372. New York: Springer.

Jurafsky, D., & Martin, J.H. 2000. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kintsch, W. 2001. Predication. *Cognitive Science* 25:173-202.

Landauer, T.K., & Dumais, S.T. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the 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.

McNamara, D.S. 2004. SERT: Self-explanation reading training. *Discourse Processes* 38:1-30.

Olde, B. A., Franceschetti, D.R., Karnavat, Graesser, A. C. & the TRG 2002. The right stuff: Do you need to sanitize your corpus when using latent semantic analysis? In W. Gray and C. Schunn Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 708-713. Mahwah, NJ: Erlbaum.

Penumatsa, P., Ventura, M., Graesser, A.C., Franceschetti, D.R., Louwerse, M., Hu, X., Cai, Z., & the TRG 2004. The right threshold value: What is the right threshold of cosine measure when using latent semantic analysis for evaluating student answers? *International Journal of Artificial Intelligence Tools.* Forthcoming.

Person, N. K., Graesser, A. C., & the TRG 2002. Human or computer? AutoTutor in a bystander Turing test. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems,* 821-830. Berlin: Springer.

VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., & Rosé, C.P. 2004. Natural language tutoring: A comparison of human tutors, computer tutors and text. Submitted to *Cognitive Science.*

VanLehn, K., Jordan, P., Rosé, C. P., Bhembe, D., Bottner, M., Gaydos, A., et al. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S.A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring,* 158-167. Berlin: Springer – Verlag.

Ventura, M., Hu, X., Graesser, A., Louwerse, M., & Olney, A. 2004. The context dependent sentence abstraction model. In K. D. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26rd Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum. Forthcoming.

Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S.P. Lajoie and M. Vivet, *Artificial Intelligence in Education,* 535-542. Amsterdam: IOS Press.