The Foundations and Architecture of Autotutor

Peter Wiemer-Hastings[1,2], Arthur C. Graesser[1,2], Derek Harter[2], and the Tutoring Research Group*[1,2,3]

[1] Department of Psychology
[2] Department of Mathematical Sciences
[3] College of Education
The University of Memphis
Memphis TN 38152-6400
{pwmrhstn, a-graesser, dharter}@memphis.edu

Abstract. The Tutoring Research Group at the University of Memphis is developing an intelligent tutoring system which takes advantages of recent technological advances in the areas of semantic processing of natural language, world knowledge representation, multimedia interfaces, and fuzzy descriptions. The tutoring interaction is based on in-depth studies of human tutors, both skilled and unskilled. Latent semantic analysis will be used to semantically process and provide a representation for the student's contributions. Fuzzy production rules select appropriate topics and tutor dialogue moves from a rich curriculum script. The production rules will implement a variety of different tutoring styles, from a basic untrained tutor to one which uses sophisticated pedagogical strategies. The tutor will be evaluated on the naturalness of its interaction, with Turing-style tests, by comparing different tutoring styles, and by judging learning outcomes.

1 Introduction

At the University of Memphis, our team of researchers from Psychology, Computer Science, and Education has begun development of an intelligent tutoring system that is fundamentally different from previous ITS's. We are taking advantage of recent technological developments as well as advances in our understanding of human tutoring protocols in order to make a tutor (called Autotutor) that can interactively communicate with a student in natural language, and produce a wide range of appropriate responses. This paper describes the motivation and foundations of this project, the basic architecture that is being developed and types of behaviors that it will support, the methods we will use for evaluating the system, and the future directions for research and development.

* The Tutoring Research Group includes Stan Franklin, Max Garzon, Barry Gholson, Doug Hacker, Xiangen Hu, Roger Kreuz, Bill Marks, Natalie Person, Fergus Nolan, Ashraf Anwar, Myles Bogner, Derek Harter, Lee McCauley, Zhaohua Zhang, Jim Hoeffner, Bianca Klettke, Kristen Link, Brent Olde, Victoria Pomeroy, Katja Wiemer-Hastings, Holly Yetman, Scotty Craig, Patrick Chipman, Melissa Ring, Charles Webster and the authors. This project is supported by grant number SBR9720314 from the National Science Foundation's Learning and Intelligent Systems Unit.

2 Motivation

Researchers have long been attempting to develop a computer tutor that can interact naturally with students to help them understand a particular subject domain. Unfortunately, however, language and discourse have constituted a serious barrier in these efforts. Language and discourse facilities have been either nonexistent or extremely limited in the most impressive and successful intelligent tutoring systems available, such as Anderson's tutors for geometry, algebra, and computer languages [1], Van Lehn's tutor for basic mathematics [26], and Lesgold's tutor for diagnosing and repairing electronic equipment [17]. There have been some attempts to augment ITS's with language and dialog facilities [13]. But such attempts have been limited by three major obstacles: (1) the inherent difficulty of getting a computer to "comprehend" the language of users, including utterances that are not well-formed syntactically and semantically, (2) the difficulty of getting computers to effectively use a large body of open-ended, fragmentary, and unstructured world knowledge, and (3) the lack of research on human tutorial dialogue.

Advances in research during the last five years make it much more feasible to develop a computer tutor which tackles the above three barriers. Our Autotutor system will "comprehend" text that the learner types into the keyboard (i.e., the initial version of our tutor will not support spoken input from the student). It will generate discourse contributions in the form of printed text, synthesized speech, graphic displays, animation, and simulated facial movements and expressions [18, 21]. The primary technological contribution of this research, however, lies in formulating helpful discourse contributions based on an analysis of human – human tutoring sessions.

3 Advances facilitating ITS development

As mentioned above, three major barriers (natural language, world knowledge, and tutorial dialog) have prevented ITS researchers from implementing tutorial dialog facilities in natural language. Recent advances have provided approximate solutions to minimizing these barriers, so an ITS with a natural language and dialog facility is much more feasible. In sections 3.1 and 3.2, we will address the two most important advances: world knowledge representation and tutorial dialogue. Space limitations force us to only briefly mention the following technical advances which will have also have a significant affect on Autotutor's chances for success:

Natural language processing The DARPA Message Understanding initiative [7], has pushed researchers away from toy problems to dealing with real-world texts, which use a wide variety of words and include complex and often ill-formed grammatical constructions.

Multimedia The ability to fluidly present not just text, but also synthesized speech, graphic displays, simulated facial movements, and animation has moved from the state-of-the-art to the commonplace.

Synthesized speech Pitch, pause, duration, amplitude, and intonation contours are among the variety of intonation cues that signal back channel feedback, affect, and emphasis [12]. Synthesized speech will allow us to provide this type of feedback to the students.

Talking heads Several researchers have recently developed relatively realistic animated talking heads that have facial features synchronized with speech [21]. This allows us to provide facial gesture feedback as well.

Fuzzy descriptions of figures and diagrams For each picture or graph that is be presented as a teaching aid, Autotutor uses a fuzzy description to specify the set of components in the picture, properties of components, spatial relationship between components, motion depicted by arrows, and so on [25].

3.1 World Knowledge representation

The fact that world knowledge is inextricably bound to natural language comprehension is widely acknowledged in psycholinguistics, cognitive science and discourse processing [9, 15, for example], but researchers in computational linguistics have not had a satisfactory approach to handling the deep abyss of world knowledge. The traditional approach to representing world knowledge in artificial intelligence has been structured representations, such as semantic networks, conceptual graphs, and rules [10]. World knowledge is frequently open-ended, imprecise, vague, and incomplete, so simple computational procedures cannot handle the role of world knowledge in understanding language and in tutoring.

Latent Semantic Analysis (LSA) provides the critical backbone for representing world knowledge in Autotutor. LSA has recently been proposed as a statistical representation of a large body of world knowledge [8, 16, for example]. LSA capitalizes on the fact that particular words appear in particular texts; the occurrence of words in texts reflects the constraints that exist in world knowledge. A statistical method called *singular value decomposition* reduces a very high-dimensional co-occurrence matrix to K dimensions (typically, 100 to 300 dimensions). Each word, sentence, or text is represented as a weighted vector on the K dimensions. The "match" (i.e., similarity in meaning, conceptual relatedness) between two words, sentences, or texts is computed as the cosine between the two vectors, with values ranging from -1 to 1. The match between two language strings can be high even though there are few if any words in common between the two strings. LSA goes well beyond simple string matches because the meaning of a language string is determined in large part by the company (other words) that each word keeps.

The empirical success of LSA has been promising and sometimes remarkable. Landauer and Dumais [16] created an LSA space from a large subset of Grolier's Academic American Encyclopedia, and then gave it the Test of English as a Foreign Language (TOEFL) from ETS. The LSA model answered 64.4% of the questions correctly, which is essentially equivalent to the 64.5% performance for college students from non-English speaking countries. Foltz et al., [8] and Kintsch [15] report other successful applications of LSA to different tasks.

LSA plays a central role in Autotutor. The *truth* of a student's contribution is evaluated by computing the maximum cosine match between a student's contribution and the entire corpus of related texts. The *relevance* of a student's contribution is evaluated by computing its match with expected answers to a question, or expected solutions to a problem. Prior to LSA, there was no empirically defensible computation of the truth and relevance of expressions with respect to a large knowledge base that is open-ended, fragmentary, imprecise, and vague. We believe that LSA will allow us to bootstrap the ITS enterprise to accommodate natural language and dialog for the first time. Autotutor provides a research platform to allow us to test this claim.

3.2 Tutorial Dialog

Researchers in education and ITS development have identified a number of ideal tutoring strategies, such as: the Socratic method [6], modeling-scaffolding-fading [24], reciprocal training [20], anchored learning [3], and others. Researchers who have examined these tutoring strategies have frequently pointed out that tutors need extensive training on the use of these sophisticated ideal tutoring strategies. Not surprisingly, therefore, these strategies do not spontaneously emerge in the repertoire of strategies of unskilled tutors — the tutors that predominate in actual school systems [11]. Previous ITS developers have abandoned attempts to incorporate most of these ideal tutoring strategies in the tutoring systems because of the barriers of natural language and world knowledge. We will implement some of these ideal tutoring strategies in Autotutor, to the extent that they are technically feasible.

Aside from these ideal tutoring strategies, recent projects have dissected the strategies used by skilled and unskilled human tutors. In some of these studies, the tutors have been highly skilled and knowledgeable about the topic [14, for example]. Our previous work on tutorial dialog [22, for example], funded by the Office of Naval Research, has examined untrained tutors with moderate domain knowledge because these tutors are most representative of tutors in actual school systems. Even though most tutors in school systems are untrained, they are surprisingly very effective compared to teachers in normal classroom environments. One-on-one human tutoring has shown effect sizes of .4 to 2.3 standard deviation units compared to classroom teaching and other suitable controls [2]. Our detailed conversational analyses of normal tutors unveiled the characteristics of the dialog that apparently are responsible for the robust learning gains [11].

One conceivable advantage of tutoring in general is an enhanced "meeting of the minds" between student and tutor. That is, the tutor infers the idiosyncratic knowledge, bugs, and misconceptions of the student – and the student's knowledge drifts toward the tutor's knowledge base. Designers of some ITS's have implemented "student modeling," to attempt to infer the student's knowledge states [1]. Discourse theories have frequently emphasized the importance of establishing shared meanings for successful communication [5]. There is a radically different perspective on the matter of common ground and student modeling, however. Researchers have cast doubt on the possibility, the need, and the pedagogical utility of detailed student modeling [19]. Our detailed analysis of actual tutoring sessions revealed that there is a very slow convergence towards shared meanings during tutoring [11]. The gap in knowledge between the tutor and student is so wide that the two parties in the conversation frequently misunderstand each other and give each other incorrect feedback. For example, tutors normally give positive responses ("Yeah", "Uh-huh") to student contributions that are vague, incoherent or error ridden; students who are lost usually say yes or nod their heads when asked "Do you understand?" [22]. The fact that the tutor manages conversation when there is a breakdown in common ground and feedback mechanisms makes tutoring a fascinating phenomenon to study from the standpoint of theories of communication and discourse processing.

The large gulf that frequently exists between the knowledge of tutors and students gives us reason to believe that it is feasible to develop a computer tutor that parallels human tutors. A key feature of effective tutoring lies in assisting students in actively constructing subjective explanations and elaborations of the material [4]. The tutor's dialog moves in a collaborative exchange might provide effective scaffolding for a student to build such self-explanations – without the computer fully knowing what the student knows.

Autotutor will simulate dialog moves in tutorial dialog of different classes of tutors. One class is unskilled tutors, the sort of tutors that exist in real school systems. Another class will be untrained tutors who acquire more experience in tutoring; the computer tutor will augment its knowledge base by storing answers that students give to questions and solutions to problems (segregating good and bad contributions). More sophisticated classes of tutors will implement various ideal tutoring strategies (such as a Socratic tutor, modeling-scaffolding-fading, and strategic hinting).

4 Autotutor Architecture

A schematic view of the architecture of Autotutor is shown in figure 1. This section describes the major processing components and knowledge bases of the system, and its overall behavior.

4.1 Topic selection

At every stage in the tutoring session, a set of production rules controls selection of a subtopic that is appropriate to the student's needs and the teacher's goals. The subtopics come from a set of instructional materials called the curriculum script, developed by experts in education and in the subject domain.

4.2 Curriculum script

The material in a curriculum script covers one topic, for example, the internet in a computer literacy class. For each topic, there is an information delivery item to

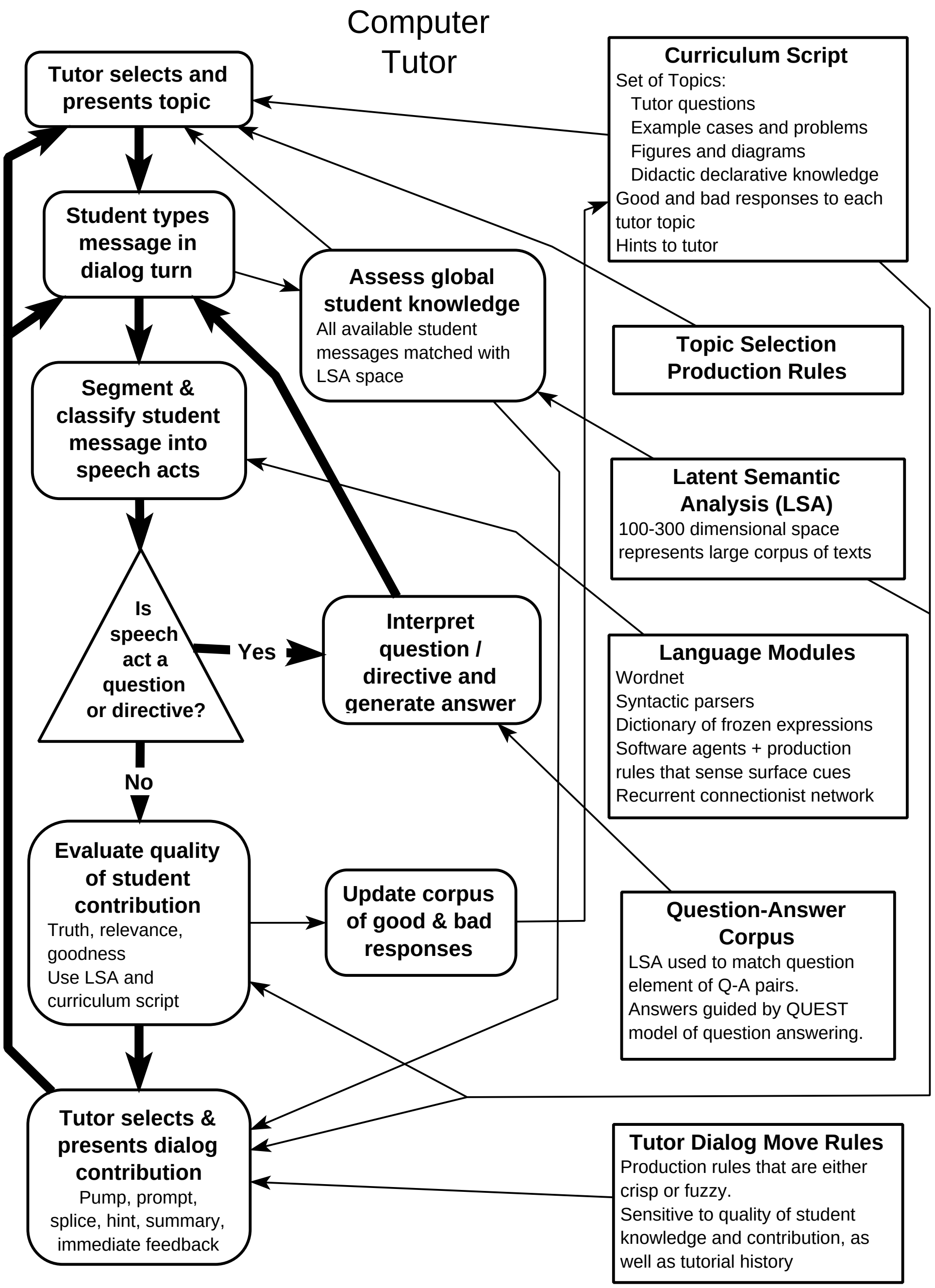
Computer Tutor
Tutor selects and presents topic
Student types message in dialog turn
Assess global student knowledge
All available student messages matched with LSA space
Segment & classify student message into speech acts
Is speech act a question or directive?
Yes
No
Interpret question / directive and generate answer
Evaluate quality of student contribution
Truth, relevance, goodness
Use LSA and curriculum script
Update corpus of good & bad responses
Tutor selects & presents dialog contribution
Pump, prompt, splice, hint, summary, immediate feedback
Curriculum Script
Set of Topics:
Tutor questions
Example cases and problems
Figures and diagrams
Didactic declarative knowledge
Good and bad responses to each tutor topic
Hints to tutor
Topic Selection Production Rules
Latent Semantic Analysis (LSA)
100-300 dimensional space represents large corpus of texts
Language Modules
Wordnet
Syntactic parsers
Dictionary of frozen expressions
Software agents + production rules that sense surface cues
Recurrent connectionist network
Question-Answer Corpus
LSA used to match question element of Q-A pairs.
Answers guided by QUEST model of question answering.
Tutor Dialog Move Rules
Production rules that are either crisp or fuzzy.
Sensitive to quality of student knowledge and contribution, as well as tutorial history

Fig. 1. The architecture of Autotutor. Rectangles indicate knowledge sources, rounded rectangles are processes, the triangle is a choice point, and the very heavy arrows indicate the general flow of control.

set up the common ground between the student and Aututotor and a seed question to "get the ball rolling." The remainder of the curriculum script consists of a set of subtopics. There are four types of subtopics: (1) simple question/answer, (2) question/answer with didactic content, (3) problem solution, and (4) picture question/answer.

Each subtopic is ranked on sophistication (high, medium, low) and on chronological order (early, middle, late). Some topics may have specific ordering constraints, for example that subtopic A must be covered *before* subtopic B. Associated with each subtopic is: an ideal complete and correct answer; a set of additional good answers (which grows with experience), ranked good, better, best; a set of bad answers (which grows with experience); a set of hints, ranked for low, medium, and high-quality student contributions (i.e. a low-ranked hint provides more basic information than a high-ranked hint); a set of question that the student would be likely to ask, with appropriate answers; and a good succinct summary.

The set of curriculum scripts provides a rich set of responses from which Autotutor can choose, based on the evaluation of the student's contribution (described below) and on the dialogue selection rules (described in section 4.4).

4.3 Evaluating the quality of student contributions

Human tutors are sensitive to the overall quality of student contributions during the collaborative process of answering a question or solving a problem. This component evaluates the quality of the students' speech acts that are classified as Contributions, but not Questions and Directives. As mentioned above, latent semantic analysis will be used in these assessments of the contributions' truth, relevance, and quality.

4.4 Tutor dialogue moves

Our analysis of untrained tutors uncovered a set of dialogue moves that are triggered under specific conditions during the collaborative evolution of an answer to a question or a solution to a problem [11]. Some of these moves are specified below:

Pumping: The tutor pumps the student for more information during the early stages of answering a particular question (or solving a problem).

Prompting: Tutors supply the student with a discourse context and prompt them to fill in a missing word, phrase, or sentence.

Immediate feedback: Tutors are normally polite conversation partners, so they are reluctant to give negative feedback after student contributions that have poor quality [23]. Instead, they give positive, neutral, or indirect feedback.

Splicing: A tutor jumps in and splices correct information as soon as the student produces a contribution that is obviously error-ridden.

Hinting: When the student is having problems answering a question or solving a problem, the tutor gives hints by presenting a fact, asking a leading question, or reframing the problem.

Summarizing: Unskilled tutors normally give a summary that recaps an answer to a question or solution to a problem.

Autotutor will implement these dialogue moves to the extent possible using the dialogue move production rules described below.

4.5 Dialogue move production rules

Autotutor includes another set of fuzzy production rules that specify which dialogue response the tutor will make after a student turn. These are conditionalized on the content of the curriculum script, the dialogue history, and the quality of the student's contribution during the last turn, the cumulative quality of the student's knowledge, and the cumulative quality of the student-tutor exchange. For example, the following simple production rule is associated with immediate positive feedback for student contribution C:

```
IF [scriptComponent = Question(j)
    AND max(similarity(C, good-answer(j)) > Threshold]
THEN [produceFeedback: "That's right."]
```

The style and expertise of the tutor is defined in part by the production rules that capture these dialogue rules and the values of the threshold parameters. The production rules will be different for the unskilled tutor with minimal experience and a tutor with much experience (i.e., a large accumulated list of good answers and bad answers) which uses frontier learning. An important goal of this project is to simulate the contributions of different classes of tutors which vary in experience and in the use of particular pedagogical strategies.

5 Evaluating Autotutor

The primary objective of the evaluation is to assess the pedagogical quality and conversational aptness of the simulated tutor contributions. A tutor contribution should have pedagogical value, be relevant to the conversational context, and be informative. While a stable prototype version of Autotutor is still being developed, informal assessments of the system will prevail. When the computer tutor approaches the arena of supplying reasonable contributions, we will make more systematic evaluations of the tutor's contributions. The following types of evaluations are planned:

Initial evaluations: College students will interact with the computerized tutor and supply a sample of tutorial dialogs. The students will enter information by keyboard and the tutor will display information on computer screen and through simulated speech during these interactions. The content of the tutor's contributions in these transcripts will be analyzed by experts in discourse analysis, experts in education, and graduate research assistants.

Turing tests: College students will read transcripts in which half of the contributions are generated by Autotutor, and half by human experts. Then they will make a decision as to whether a human or computer generated each contribution. We will perform a standard signal detection analysis that segregates a true discrimination between human and computer from response biases in making these judgments; we will collect hit rates, false alarm rates, d scores, and Ag discrimination scores.[1]

Comparison of tutors: As discussed earlier, we will compare different classes of tutors that embrace different ideal tutoring strategies. These models will be the same except for the production rules that select topics and dialogue moves. Trained judges will assess the tutor's contributions with respect to conversational smoothness and pedagogical value, as discussed above.

Learning outcomes: Although the primary objective of this research does not address how well students learn with these computerized tutors, we do plan on conducting preliminary evaluations during the third year of the project. College students in a Computer Literacy course will be randomly assigned to use one of the different classes of tutors described above, or to a reading control. The students' learning of the domain material will be tested with a variety of measures of understanding.

6 Conclusion

We are in the first year of development of the system. As of mid-April 1998, we have a running prototype which includes latent semantic analysis, the talking head, the curriculum script, and the topic and subtopic selection rules. Later versions of the system will include in-depth natural language processing and a mechanism for automatically inferring from a picture a set of fuzzy descriptions of the relationships between the objects in the picture. As mentioned above, the focus of our efforts will now shift to evaluating the technical components of the system, evaluating the naturalness of its interactions, and developing the additional, more challenging components.

References

1. J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4:167–207, 1995.
2. B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984.
3. J. D. Bransford, S. R. Goldman, and N. J. Vye. Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg and L. Okagaki, editors, *Influences on children*, pages 147–180. Erlbaum, Hillsdale, NJ, 1991.

[1] An alternative method of performing a Turing test would be to have a human expert provide contributions on-line during the tutorial sessions at random points in the dialogue; the computer would supply contributions at the remaining points.

4. M. T. H. Chi, N. de Leeuw, M. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18:439–477, 1994.
5. H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.
6. A. Collins. Teaching reasoning skills. In S. Chipman, J. Segal, and R. Glaser, editors, *Thinking and learning skills*, pages 579–586. Erlbaum, Hillsdale, NJ, 1985.
7. DARPA. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufman Publishers, San Francisco, 1995.
8. P. Foltz. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers*, 28:197–202, 1996.
9. M. A. Gernsbacher, editor. *Handbook of Psycholinguistics.* Academic Press, San Diego, CA, 1994.
10. A. Graesser and L. Clark. *Structures and procedures of implicit knowledge.* Ablex, Norwood, NJ, 1985.
11. A. C. Graesser, N. K. Person, and J. P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:359–387, 1995.
12. J. Hirschberg and G. Ward. The interpretation of the high-rise question contour in english. *Journal of Pragmatics*, 24:407–412, 1995.
13. V. Holland, J. Kaplan, and M. Sams. *Intelligent language tutors.* Erlbaum, Mahwah, NJ, 1995.
14. G. D. Hume, J. Michael, A. Rovick, and M. W. Evens. Hinting as a tactic in one-on-one tutoring. *The Journal of the Learning Sciences*, 5:23–47, 1996.
15. W. Kintsch. *Comprehension: A paradigm for cognition.* Cambridge University Press, Cambridge, MA, inpress.
16. T. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
17. A. Lesgold, S. Lajoie, M. Bunzo, and G. Eggan. Sherlock: A coached practice environment for an electronics troubleshooting job. In J. H. Larkin and R. W. Chabay, editors, *Computer-assisted instruction and intelligent tutoring systems*, pages 201–238. Erlbaum, Hillsdale, NJ, 1992.
18. D. Massaro and M. Cohen. Perceiving talking faces. *Psychological Science*, 4:104–109, 1995.
19. D. Newman. Is a student model necessary? Apprenticeship as a model for ITS. In D. Bierman, J. Breuker, and J. Sandberg, editors, *Artificial intelligence and education.* IOS, Amsterdam, 1989.
20. A. S. Palinscar and A. Brown. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1:117–175, 1984.
21. C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20:1–46, 1996.
22. N. K. Person, A. C. Graesser, J. P. Magliano, and R. J. Kreuz. Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences*, 6:205–229, 1994.
23. N. K. Person, R. J. Kreuz, R. Zwaan, and A. C. Graesser. Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, 13:161–188, 1995.
24. B. Rogoff. *Apprenticeship in thinking.* Oxford University Press, New York, 1990.
25. B. Tversky and K. Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15:121–149, 1983.
26. K. Van Lehn. *Mind bugs: The origins of procedural misconceptions.* MIT Press, Cambridge, MA, 1990.