Automatic Textual Feedback for Guided Inquiry Learning

Steven TANIMOTO, Susan HUBBARD, and William WINN
Online Learning Environments Laboratory
Box 352350, Dept. of Computer Science and Engineering
University of Washington, Seattle, WA, 98195, USA

Abstract. We briefly introduce the online learning environment INFACT, and then we describe its textual feedback system. The system automatically provides written comments to students as they work through scripted activities related to image processing. The commenting takes place in the context of an online discussion group, to which students are posting answers to questions associated with the activities. Then we describe our experience using the system with a class of university freshmen and sophomores.

1. Introduction

Timely feedback has been found in the past to improve learning [1]. However, it can be a challenge to provide such feedback in large classes or online environments where the ratio of users to teachers and administrators is high. We report here on an experimental system that provides automated feedback to students as they work on activities involving elementary image processing concepts.

1.1 Project on Intensive, Unobtrusive Assessment

The motivation for our project is to improve the quality of learning through better use of computer technology in teaching. We have focused on methods of assessment that use as their evidence not answers to multiple-choice tests but the more natural by-products of online learning such as students' user-interface event logs, newsgroup-like postings and transcripts of online dialogs. By using such evidence, students may spend more of their time engaged in the pursuit of objectives other than assessment ones: completing creative works such as computer programs and electronic art, or performing experiments using simulators in subject areas such as kinematics, chemical reactions, or electric circuits. (We currently support programming in Scheme and Python, and performing mathematical operations on digital images.)

Various artificial intelligence technologies have the potential to help us realise the goal of automatic, unobtrusive diagnostic educational assessment from evidence naturally available through online learning activities. These technologies include textual pattern matching, Bayesian inference, and Latent Semantic Indexing [4]. In this paper, we focus on our experience to date using textual pattern matching in this regard.

1.2 Facet-Based Pedagogy

Our project is studying automatic methods for educational assessment in a context in which multiple-choice tests are usually to be avoided. This means that other kinds of evidence must be available for analysis, and that such evidence must be sufficiently rich in information that useful diagnoses of learning impediments can be made. In order to obtain this quality of evidence, the learning activities in which our assessments are performed are structured according to a "facet-based pedagogy."

A *facet* is an aspect, conception, approximate state of understanding, or state of skill with regard to some concept, phenomenon, or skill. Minstrell [5] uses the term "facet" to refer to a variation of and elaboration of DiSessa's phenomenological primitive ("p-prim") [3]. We use the term "facet" in a more general sense, so as to be able to apply a general pedagogical approach to the learning not only of conceptual material such as Newton's laws of motion but also of languages and skills.

The facet-based pedagogical structure we use posits that instruction take place in units in which a cycle of teaching and learning steps proceeds. The cycle normally lasts one week. It begins with the posing of a problem (or several problems) by the instructor. Students then have one day to work on the problem individually and submit individual written analyses of the problem. Once these have been collected, students work in groups to compare and critique answers, keeping a record of their proceedings. By the end of the week, the students have to have submitted a group answer that incorporates the best of their ideas. It also must deal with any discrepancies among their individual analyses.

Students work in groups for several reasons. One is essentially social, allowing students to feel involved in a process of give-and-take and to help each other. Another is that the likely differences in students' thinking (assuming the problems are sufficiently challenging), will help them to broaden their perspectives on the issues and focus their attention on the most challenging or thought-provoking parts of the problem. And the most important reason, from the assessment point of view, to have the students work in groups is to help them communicate (to each other, primarily, as they see it, but also to us, indirectly) so as to create evidence of their cognition that we can analyse for misconceptions.

During the cycle, we expect some of the students' facets to change. The facets they have at the beginning of the unit, prior to the group discussion, are their preconceptions. Those they have at the end of the unit are their postconceptions. We want their postconceptions to be better than their preconceptions, and we want the postconceptions to be as expert-like as possible.

In order to facilitate teaching and learning with this facet-based pedagogy, we have developed a software system known as INFACT. We describe it in the next section.

2. The INFACT Online Learning Environment

Our system, called INFACT, stands for Integrated, Networked, Facet-based Assessment Capture Tool [6, 7]. INFACT catalyses facet-based teaching and learning by (a) hosting online activities, (b) providing tools for defining specific facets and organising them, (c) providing simple tools for manual facet-oriented mark-up of text and sketches, (d) providing tools for displaying evidence in multiple contexts including threads of online discussion, and timeline sequence, and (e) providing facilities for automatic analysis and automatic feedback to students. INFACT also includes several class management facilities such as automatic assignment of student to groups

based on the students' privately entered preferences (uses the Squeaky-Wheel algorithm), automatic account creation from class lists, and online standardised testing (for purposes such as comparison to the alternative means of assessment that we are exploring).

The primary source of evidence used by INFACT is a repository of evolving discussion threads called the *forum*. Most of the data in the forum is textual. However, sketches can be attached to textual postings, and user-interface log files for sessions with tools such as an image processing system known as PixelMath [8] are also linked, automatically by the system, from textual postings.

The forum serves the facet-based pedagogical cycle by mediating the instructor's challenge problem, collecting student's individual responses and hiding them until the posting deadline at which time the "curtain'' is lifted and each student can see the posts of all members of his or her group. The forum hosts the ensuing group discussions, and provides a record of it for both the students and the instructor. Any facet-oriented mark-up of the students' messages made by the instructor or teaching assistants is also stored in the forum database. In the experiments we performed with manual and automated feedback to students, we used a combination of the forum and email for the feedback.

The facet-based pedagogy described above, as adapted for INFACT, is illustrated in Figure 1. A serious practical problem with this method of teaching is that the fourth box, "Teacher's facet diagnoses," is a bottleneck. When one teacher has to read all the discussions and interact with a majority of the students in a real class, most teachers find it impossible to keep up; there may be 25 or more students in a class, and teachers have other responsibilities than simply doing facet diagnoses. This strongly suggests that automation of this function be attempted.

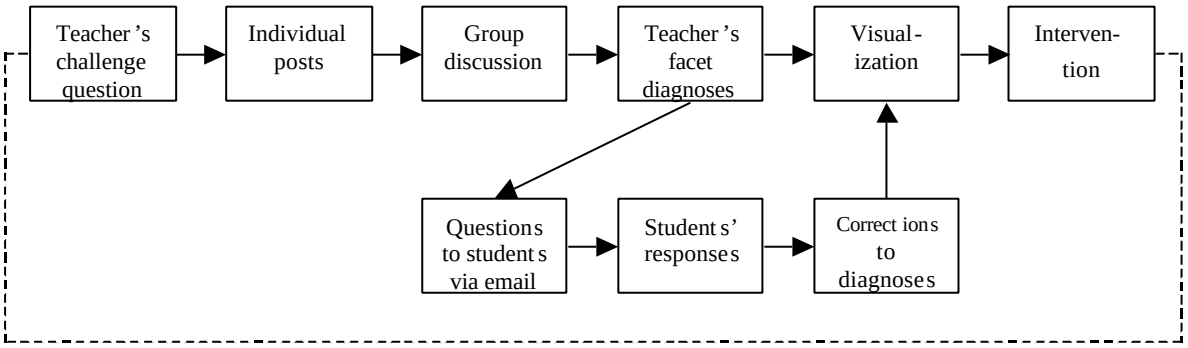

Figure 1. The INFACT pedagogical cycle. The period of the cycle is normally 1 week.

INFACT provides an interface for teachers to analyse student messages and student drawing, and create assessment records for the database and feedback for the students. Figure 2 illustrates this interface, selected for sketch-assessment mode. The teacher expresses an assessment for a piece of evidence by highlighting the most salient parts of the evidence for the diagnosis, and then selecting from the facet catalog the facet that best describes the student's apparent state of learning with regard to the current concept or capability.

In order to provide a user-customisable text-analysis facility for automatic diagnosis and feedback, we designed and implemented a software component that we call the INFACT rule system. It consists of a rule language, a rule editor, and a rule applier. The rule language is based on regular expressions with an additional construct to make it work in INFACT. The rule editor is a Java applet that helps assure that rules entered into the rule system are properly structured and

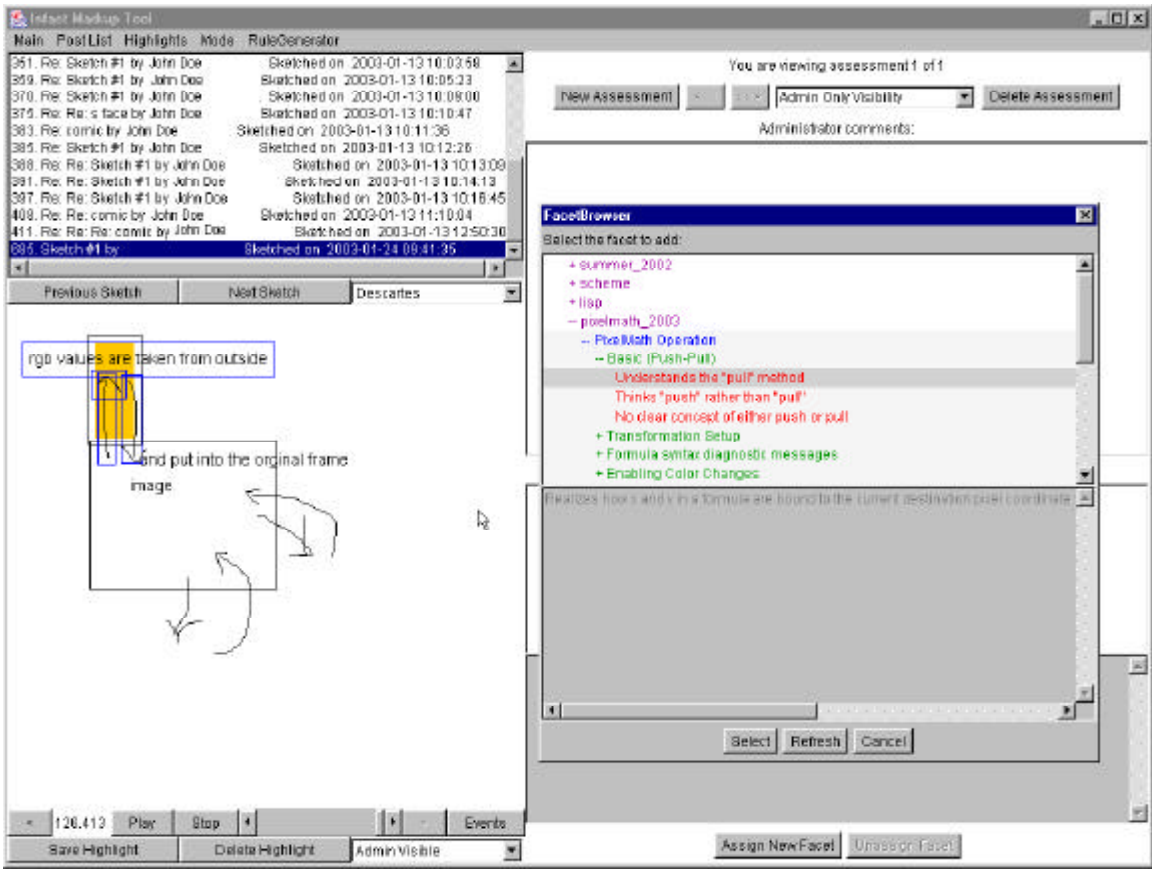

Figure 2. The manual mark-up tool for facet-based instruction. It is shown here in sketch-assessment mode, rather than text assessment mode.

written. The rule applier consists of a combination of back-end Perl scripts and a Java graphical user interface.

The INFACT rule language is based on regular expressions. These regular expressions are applied by the rule applier to particular components of text messages stored in INFACT-forum. In addition to the regular expressions, rule patterns contain "field specifiers." A field specifier identifies a particular component of a message: sender name, date and time, subject heading, body. Each instance of a field specifier will have its own regular expression. Someone creating a rule (e.g., a teacher or educational technology specialist) composes a rule pattern by creating any number of field specifier instances and supplying a regular expression for each one. Each field specifier instance and regular expression represent a subcondition for the rule, all of which must match for the rule to fire. It is allowed to have multiple instances of the same field specifier in a pattern. Therefore INFACT rules generalise standard regular expressions by making conjunction available.

The rule applier can be controlled from a graphical user interface, and this is particularly useful when developing an assessment rule base. While regular expressions are a fundamental concept in computer science and are considered to be conceptually elementary, designing regular expressions to analyse text is a difficult and error-prone task, because of the complexity of natural language, particularly in the possibly broken forms typically used by students in online writing. Therefore we designed the rule applier to make as easy as possible to test new rules. Although a complete rule specifies not only a condition, but also an action, the rule applier can be used in a way that safely tests conditions only. One can easily imagine that if it didn't have this facility, a teacher testing rules in a live forum might create confusion when the rules being debugged cause email or

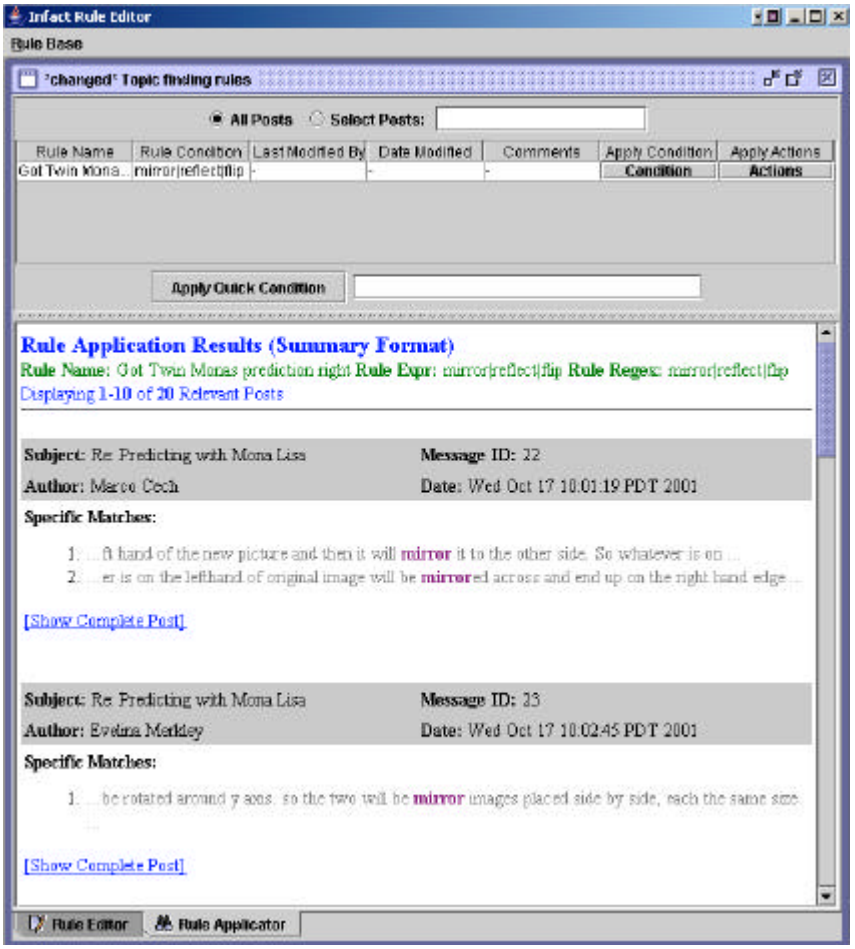

Figure 3. The "hit list" returned by the rule applier in testing mode.

INFACT postings to be sent to students inappropriately. When applying rules in this safe testing mode, the rule actions are not performed, and the results of condition matching are displayed in a "hit list" not unlike the page of hits returned by a search engine such as Google. This is illustrated in Figure 3. It is also possible to learn rules automatically [2], but this study did not use that facility.

3. The Study

The automated feedback system was tested in a freshman class for six weeks out of a ten-week quarter. The class was given in a small computer lab where each student had their own machine. Eighteen students completed the course and provided usable data. They were randomly divided into three groups, Arp, Botero and Calder. Almost all of the work discussed here was done collaboratively within these groups.

In addition to testing the usability and reliability of the automatic feedback system for instruction, the class was used to conduct a simple study in which the effectiveness of the automatic system was compared with the effectiveness of feedback provided by an instructor. A "no-feedback" condition served as a control. The three feedback conditions were rotated through the three groups using a within-subjects design so that every student had each kind of feedback for two weeks over the six-week period. The feedback study began with the fourth week of class. The order of the types of feedback was different for each group. Each two-week period required the students to complete exercises in class and as homework. Every week, activities were

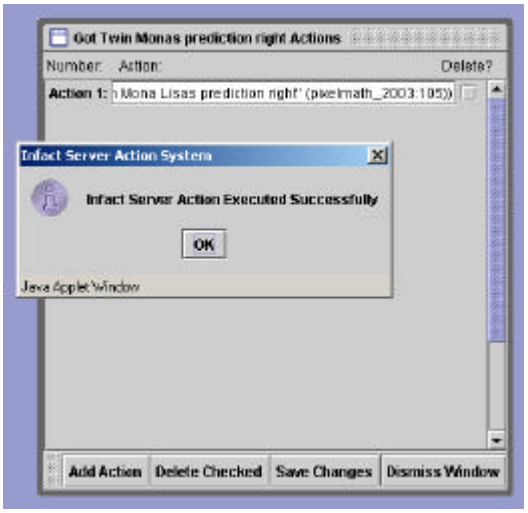

Figure 4. Feedback to the teacher/administrator from the action subsystem of the rule system.

assigned requiring each student to find the solution to a problem set by the instructor (a PixelMath formula, a strategy, some lines of Scheme code) and to post that solution to INFACT Forum by mid-week. The group then had the rest of the week to come to a consensus on the solution and to post it. At the end of the two-weeks, before the groups rotated to the next type of feedback, students took a short on-line post-test over the content covered in the preceding two weeks.

The automatic feedback was provided in the manner described above. The human feedback was provided by an instructor ("Alan"). During the class, Alan sat at one of the lab computers watching posts come into INFACT Forum from the group whose turn it was to receive human feedback. As each post arrived, he responded. Out of class, Alan checked the forum every day and responded to every post from the designated group. Students in the no-feedback group were left to their own devices.

A number of data sources were available. These included the scores on the post-tests, the content of the students' posts and the feedback provided automatically and by Alan, interviews with selected students at the end of each two-week period conducted by a research assistant, questionnaires, and observations of the class by three research assistants. The class instructor and Alan were also interviewed.

4. Findings

Analysis of the post-test scores showed no statistically reliable differences among the groups as a function of the type of feedback they received, nor significant interactions among group, feedback, or the order in which the groups received feedback. There are two explanations for this finding, aside from taking it as evidence that the automatically-provided feedback was neither more nor less effective than that provided by Alan, and that neither was better than no feedback. First, the small number of students in each group reduced the statistical power of the analysis to the point where type-two errors were a real possibility. Second, the first no-feedback group was quick to organize itself and to provide mutually-supporting feedback within its members. This proved to be extremely effective for this group (Arp) and subsequently also for Botero and Calder when it was their turn not to receive feedback.

However, examination of other data sources showed some differences between the automatic and Alan's feedback, as well as some similarities. First, both encountered technical problems. For the first few sessions, the automatic feedback system was not working properly.

This made it necessary for a research assistant to monitor the posts from the automatic feedback group and to decide from the rules which prepared feedback statement to send. Fortunately, the bug was fixed and the Wizard-of-Oz strategy was quickly set aside. Also, Alan soon discovered that posting his feedback to INFACT Forum took too long as the system acted sluggishly. It was therefore decided to send the "human" feedback to the students' personal email accounts. This was much quicker. However, it required the students to have their email programs open at the same time as INFACT Forum and PixelMath. With so many windows open, some students did not notice Alan's feedback until some time after it had been sent. Some even minimized their email windows to make their screens more manageable and did not read the feedback until some time after it was sent, if at all.

The most obvious difference between the automatic and the human feedback was that the automatic feedback was very quick, while it took Alan time to read students' posts, consider what to reply, to type it and send it. This delay caused some minor frustration. One observer reported students posting to INFACT and then waiting for Alan's response before doing anything else. Several students were even seen to turn in their seats and watch Alan from behind while they were waiting for feedback. Also, out of class, Alan's feedback was not immediate, as he only checked the forum once a day. Automatic feedback was provided whenever a student posted something, whether during class or out of class.

Next, the automatic feedback responses were longer and more detailed than Alan's. This was because they had been generated, with careful thought, ahead of time, while Alan responded on the fly. Alan also mentioned that he often had difficulty keeping up with the student posts during class and that he had to be brief in order to reply to them all.

Over the six weeks Alan posted close to 300 messages. The automatic system sent less than 200. The main reason for this difference seems to be Alan's tendency to respond in a manner that encouraged the development of discussion threads. While both types of feedback asked questions of students and asked them to post another message as a matter of course ("Why do you think that is?", "Try again and post your response."), this tactic produced only one follow-on post to an automatic feedback message during the six weeks of the study.

Though posting shorter messages, Alan was better than the automatic system at deciding what a student's particular difficulty might be, and responding more flexibly and particularly to individual students' posts. Some of the students said they preferred Alan's feedback for this reason, finding the automatic feedback too general or less directly relevant to their particular difficulties or successes. Moreover, Alan could sometimes determine more precisely than the automatic system what was causing a student to have a problem. In such cases, he would often suggest a strategy for the student to try, rather than giving direct feedback about the student's post. Alan also referred students to other students' posts as part of his feedback. Because he was monitoring all of the posts from the group, while the students themselves might not be, he knew if another student had solved a problem or come up with a suggestion that would be useful to the student to whom he was currently responding, and did not hesitate to have the student look at the other's post. This also speeded up the feedback process somewhat. On two occasions, Alan was able to spot common problems that were then addressed for everyone in the next class session.

The students found Alan's feedback more personal. He made typos and used incomplete sentences. The automatic system did not. He used more vernacular and his posts reflected a more friendly tone. Alan also made an occasional mistake in the information he provided through feedback, though, fortunately, these were quickly identified and put right. In spite of this, most students preferred interacting with a human rather than the automatic system.

Finally, as we mentioned above, the first group to receive no feedback, Arp, compensated for this by providing feedback and other support to each other. By coincidence, students in Arp, more than in Botero and Calder, had, by the fourth week, developed the habit of helping each other through the forum. It turns out that Arp also contained the strongest students in the class who, collectively, had strength in all the skills required in the course. As a result, requests for help from one group member were answered without fail, in one case by ten responses from the other group members. One result of this was that, when it was Arp's turn to receive the system's feedback and then Alan's, they had come to rely on it. (The students who stopped work until Alan replied to their posts, whom we mentioned above, were all from Arp.)

To summarize, the automatic feedback system worked. Initial technical problems were quickly solved and the students received detailed and mostly relevant feedback on their posts to INFACT Forum. The comparison to human feedback points to improvements that should be considered. First, it would be useful if the system could cross-reference student posts so that students could be referred to each other's contributions in a way that proved effective in Alan's feedback. More generally, the ability of feedback from the automatic system to generate more collaboration among the students would be an important improvement. Second, the ability of the system to better diagnose from posts the reasons students were having problems would be useful. This would allow the system to sustain inquiry learning for more "turns" in the forum, rather than giving the answer, or suggesting a particular strategy to try. Finally, any changes that made the automatic system appear to be more human would make it better received by students.

Acknowledgments

The authors wish to thank E. Hunt, R. Adams, C. Atman, A. Carlson, A. Thissen, N. Benson, S. Batura. J. Husted, J. Larsson, D. Akers for their contributions to the project and the National Science Foundation for its support under grant EIA-0121345.

References

[1] Black, P., and Wiliams, D. 2001. Inside the black box: Raising standards through classroom assessment. Kings College London Schl. of Educ. http://www.kcl.ac.uk/depsta/education/publications/Black%20Box.pdf.

[2] Carlson, A., and Tanimoto, S. 2003. Learning to identify student preconceptions from text, *Proc. HLT/NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing*.

[3] diSessa, A. 1993. Toward an epistemology of physics. *Cognition and Instruction*, 10, 2&3, pp.105-225.

[4] Graesser, A.C., Person, N., Harter, D., and The Tutoring Research Group. 2001a. Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*.

[5] Minstrell, J. 1992. Facets of students' knowledge and relevant instruction. In Duit, R., Goldberg, F., and Niedderer, H. (eds.), *Research in Physics Learning: Theoretical Issues and Empirical Studies*. Kiel, Germany: Kiel University, Institute for Science Education.

[6] Tanimoto, S. L., Carlson, A., Hunt, E., Madigan, D., and Minstrell, J. 2000. Computer support for unobtrusive assessment of conceptual knowledge as evidenced by newsgroup postings. *Proc. ED-MEDIA 2000*, Montreal, Canada, June.

[7] Tanimoto, S., Carlson, A., Husted, J., Hunt, E., Larsson, J., Madigan, D., and Minstrell, J. 2002. Text forum features for small group discussions with facet-based pedagogy, *Proc. CSCL 2002*, Boulder, CO.

[8] Winn, W., and Tanimoto, S. 2003. On-going unobtrusive assessment of students learning in complex computer-supported environments. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago IL.