Pilot-Testing a Tutorial Dialogue System That Supports Self-Explanation

Vincent Aleven, Octav Popescu, and Kenneth Koedinger

Human Computer Interaction Institute
Carnegie Mellon University
{aleven, koedinger}@cs.cmu.edu, octav@cmu.edu

Abstract. Previous studies have shown that self-explanation is an effective metacognitive strategy and can be supported effectively by intelligent tutoring systems. It is plausible however that students may learn even more effectively when stating explanations in their own words and when receiving tutoring focused on their explanations. We are developing the Geometry Explanation Tutor in order to test this hypothesis. This system helps students, through a restricted form of dialogue, to construct general explanations of problem-solving steps in their own words. We conducted a pilot study in which the tutor was used for two class periods in a junior high school. The data from this study suggest that the techniques that we chose to implement the dialogue system, namely a knowledge-based approach to natural language understanding and classification of student explanations, are up to the task. There are a number of ways in which the system could be improved within the current architecture.

1 Introduction

Recently, many researchers have embraced the notion that tutorial dialogue systems will make a dramatically more effective "3rd generation" of computer-based instructional systems (Graesser, et al., 2001; Evens, et al., 2001; Rosé & Freedman, 2000; Aleven, 2001). But what pedagogical approaches should underlie the dialogues conducted by these systems? A number of cognitive science studies have shown that self-explanation is an effective metacognitive strategy (Bielaczyc, Pirolli, & Brown, 1995; Chi, 2000; Renkl, et al., 1998). That is, when students study textbooks or worked-out examples, they learn with greater understanding to the extent that they explain the materials to themselves. However, not all students self-explain spontaneously and even when prompted to self-explain, it is difficult for students to arrive at good explanations (Renkl, et al., 1998).

This has led a number of researchers to investigate how self-explanation can be supported effectively by intelligent tutoring systems (Aleven & Koedinger, in press; Conati & VanLehn, 2000) or other instructional software (Renkl, in press). In previous work, we showed that even simple means of supporting self-explanation within an intelligent tutoring system, such as menus, can help students learn with greater understanding, as compared to tutored problem solving without self explanation (Aleven & Koedinger, in press). It is plausible that students learn even better when they explain

S.A. Cerri, G. Gouardères, and F. Paraguaçu (Eds.): ITS 2002, LNCS 2363, pp. 344–354, 2002.
© Springer-Verlag Berlin Heidelberg 2002

in their own words. However, the potential advantages of "free-form" explanation do not seem to materialize when the system merely prompts students to explain but does not provide feedback on students' explanations (Aleven, et al., 2000).

Our long-term goal is to find out whether a tutorial dialogue system can effectively tutor students as they produce self-explanations and whether it thereby can help them learn with greater understanding. We are developing a tutorial dialogue system that supports self-explanation in the domain of geometry, the Geometry Explanation Tutor (Aleven, Popescu, & Koedinger, 2001). This system helps students, through a restricted form of dialogue, to produce explanations that not only get at the right mathematical idea but also state the idea with sufficient precision. The system evaluates whether students' explanations are correct and sufficiently precise and is able to detect common omissions and errors in explanations. We have opted for a knowledge-based approach to natural language understanding, with a logic-based representation of semantics, similar in spirit to those discussed in Allen (1995). We have presented our arguments for our choice elsewhere (Popescu & Koedinger, 2000). Further, we have opted to keep dialogue management as simple as possible, adopting a "submit and critique" approach described below.

In this paper we present the results of a short pilot study in a school and discuss how far along we are towards having a robust and effective tutorial dialogue system that is ready for the classroom. Further, we discuss whether the techniques we have chosen to implement our tutorial dialogue system are adequate and were a good choice.

2 The Geometry Explanation Tutor

The Geometry Explanation Tutor was built on top of an existing Cognitive Tutor (Anderson, et al., 1995) for geometry problem solving, the Geometry Cognitive Tutor™. This tutor was developed by our research group in tandem with a full-year high school geometry curriculum. The combination of these two components has been shown to be better than traditional geometry classroom instruction (Koedinger, et al, 2000). The tutor and the curriculum are being marketed commercially and are in use in about 100 schools in the United States (see http://www.carnegielearning.com).

The Geometry Explanation Tutor provides for guided problem-solving practice. In addition, it requires that students provide general explanations of their problem-solving steps in their own words (Aleven, et al., 2001). It helps students, through a restricted form of dialogue, to improve their explanations and arrive at explanations that are mathematically precise (see Figure 1, which shows a dialogue with the system collected during the pilot study described below). So far, the dialogues deal with the topics covered in one of the six units that make up the tutor curriculum, the Angles unit, which deals with the geometric properties of angles.

The system's architecture has been described elsewhere (Aleven, et al, 2001) so here we provide only a short outline. An important knowledge source is the system's hierarchy of 149 explanation categories, which represent the most common ways in which students express or attempt to express geometry rules in their own words. For

each relevant geometry rule, the hierarchy contains one or more categories representing correct and complete ways of stating the rule. For example, category COMPLEMENTARY-ANGLES-SUM-90 represents all correct and complete statements of the definition of complementary angles, including "the sum of the measures of complementary angles is 90 degrees". For each relevant geometry rule the explanation hierarchy also contains numerous categories that represent commonly-occurring incomplete or incorrect ways of stating the rule. For example, category COMPLEMENTARY-ANGLES-90 represents sentences meaning "complementary angles are 90 degrees," which falls just short of being a complete and correct explanation.

Student-Tutor Dialogue	**Explanation Categories**
1. **Student:** whole and parts postulate **Tutor:** The Geometry Tutor is not able to understand your explanation. Sorry! Please add to your explanation or type something different.	None.
2. **Student:** two complementary angles equal 90 degrees **Tutor:** You are going in the right direction. Exactly what kinds of angles are 90 degrees?	EQUAL, **ANGLES-90**, COMPLEMENTARY-ANGLES-90
3. **Student:** two adjacent complementary angles equal 90 degrees **Tutor:** Correct, except for one thing: what exactly is 90?	**COMPLEMENTARY-ANGLES-90**, ADJACENT-ANGLES, EQUAL
4. **Student:** two adjacent complementary angles equal nine degrees **Tutor:** That may not be the right number. Can you state a geometry rule that justifies your answer?	COMPLEMENTARY-ANGLES, EQUAL, ADJACENT-ANGLES, **QUANTITY**
5. **Student:** two adjacent complementary angles equal 90 degrees **Tutor:** Correct, except for one thing: what exactly is 90?	**COMPLEMENTARY-ANGLES-90**, EQUAL, ADJACENT-ANGLES
6. **Student:** two adjacent complementary angles add up to 90 degrees **Tutor:** (Implicit feedback saying that the explanation is correct.)	**COMPLEMENTARY-ANGLES-SUM-90**, ADJACENT-ANGLES-SUM

Fig. 1. Example dialogue with the Geometry Explanation Tutor

The system has a natural language understanding (NLU) component, primarily knowledge-based, whose task is to construct a logic-based representation of the semantic content of student input and classify that representation with respect to the explanation categories. In this regard our system is different from many other tutorial dialogue systems, which rely on statistical approaches (Graesser, et al., 2001), keyword spotting (Evens et al., 2001), or bypass language understanding altogether (Heffernan & Koedinger, 2000) and from systems that use hybrid approaches combining deep and shallow methods (e.g., Wahlster, 2001).The NLU component uses a left-corner chart parser with a unification-based grammar formalism to parse the input (Rosé & Lavie, 1999). It uses the Loom description logic system (MacGregor, 1991) to construct a semantic representation of student input. Once the semantic representa-

tion has been constructed, it is classified with respect to the explanation hierarchy. This may result in a set of explanation categories, as is illustrated in Figure 1, rightmost column.

If the student's explanation was classified as a complete and correct statement of an applicable geometry theorem, the tutor accepts the explanation, as illustrated in step 6 of the dialogue shown in Figure 1. Otherwise, if the explanation was classified under one or more categories that represent incomplete or incorrect statements of an applicable geometry theorem, the tutor selects one of these categories randomly and displays the feedback message associated with the selected category. In Figure 1, the explanation category on which the tutor feedback was based is shown in bold face in the rightmost column. The system also appropriately handles explanations that are merely references to geometry rules (student gave the name of a geometry rule) and explanations that focus on the wrong rule. Since the pilot study described in this paper, a number of improvements have been made to the system, as described below. Nonetheless, it is fair to say that the techniques for dialogue management used by the Geometry Explanation Tutor are fairly straightforward compared to those used in other systems (Graesser, et al., 2001; Evens, et al., 2001; Wahlster, 2001).

3 A Pilot Study

At the end of the 2000-2001 school year, we conducted a pilot study in order to get a sense of how well the system was working. During this study, the Geometry Explanation Tutor was used briefly in a suburban junior high school in the Pittsburgh area, as part of a 9th-grade Integrated Mathematics II course. This course covered a number of topics in geometry and algebra. Approximately 30 students of ages 14 and 15 participated in the pilot study. The students were "honors students," which means that within their school they were among the best of their grade level in terms of academic ability and diligence. During two 40-minute class periods, the students worked in pairs on the Geometry Explanation Tutor, running the tutor on the school's wireless PC laptops. Earlier during the semester, they had learned about angles and their interrelationships (the topics covered in the tutor's Angles unit, on which we focused in the study) but they did not have any computer tutoring related to these topics.

4 Effectiveness of Student-Tutor Dialogues

The logs of the student-tutor interactions were analyzed in order to evaluate how effective the student-system dialogues were. The logs contained information about 185 dialogues comprising 791 explanation attempts to explain a geometry theorem or definition, or 12.3 ± 4.6 dialogues per pair of students. Students arrived at a complete explanation in 75% of the 185 dialogues. About half of the incomplete dialogues occurred simply because the bell rang at the end of the period. For the other half, the logs indicate that the students' work on the problem ended abnormally. Such abnormal

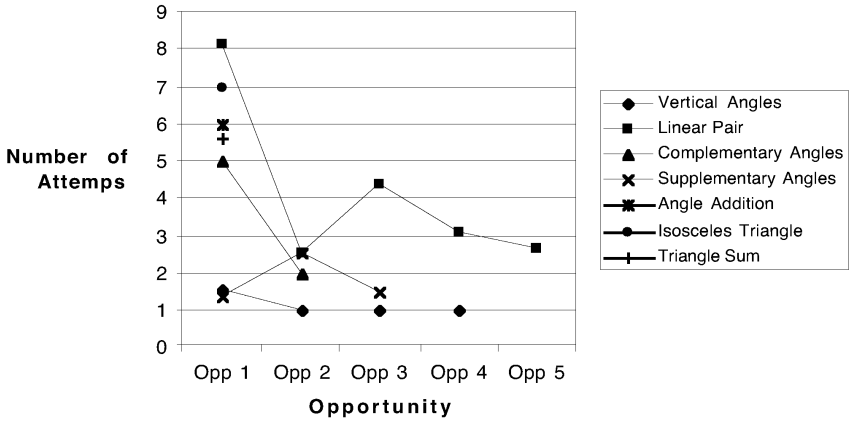

Fig. 2. Number of attempts to explain a given geometry rule, by opportunity

endings were especially likely to occur with geometry theorems that require longer statements (angle addition and angle bisection) and seem to have been caused by long system response times.

First, we looked at the number of attempts that it took students to complete their explanations. This variable provides a measure of how difficult it is to explain problem-solving steps, assisted by the system's feedback. In advance, we did not have a clear expectation of what the value of this variable should be. However, if practice with the system helps students learn to explain, one should see a decline in the number of attempts needed to explain any given theorem as students gain experience.

Overall, it took students 3.6 ± 4.5 attempts to get an explanation right. This average breaks down as follows: On the first opportunity to explain any given theorem, students needed 4.7 ± 5.6 attempts, whereas on later opportunities, they needed 2.5 ± 2.5 attempts. This decrease over time in the number of attempts was observed for most of the theorems, as is illustrated in Figure 3, which provides a more detailed break-down of the data. This figure shows also that some theorems are considerably more difficult to explain than others. Overall, the number of attempts that were needed to arrive at complete explanations seems reasonable. Further, our expectation that this number would go down as students gained experience was clearly borne out. Thus, students learned to explain the various geometry rules as they worked with the tutor.

A different way of measuring the effectiveness of the student-system dialogues is to see how often students were able to improve their explanation from one attempt to the next, assisted by the system's feedback. *A priori,* it is not clear on what percentage of attempts one would expect to see such progress. If students would rarely make progress, this would obviously not be a good sign. On the other hand, if they would make progress on every attempt, this might indicate that the tutor feedback makes the task too easy for students. As a very rough rule of thumb, then, let us say that a minimum criterion is that students make progress more often than not.

We define progress as follows: an attempt at explaining a geometry rule constitutes progress over a previous attempt if

(a) it is a correct and complete statement of an applicable geometry rule, or
(b) if it is closer to a complete statement than the previous attempt, meaning that it classifies under a more specific category in the explanation hierarchy, or
(c) if the attempt focuses on an applicable geometry rule whereas the previous attempt focused on a geometry rule that does not apply to the current step.

For example, in the dialogue shown in Figure 1, attempts 2, 5, and 6 constitute progress according to this criterion. The actual criterion is slightly more complicated due to the fact that some explanations are references to geometry rules and also because explanations can fall under multiple categories.

We define regression as the opposite of progress, see for example step 4 in Figure 1. Some explanations constitute neither progress nor regression. These may be explanations that classify under the same set of categories as the previous attempt or explanations that are classified under a different set of categories but constitute "lateral movement". This means either that the explanation is better than the previous attempt in one respect but worse in another, or that the explanation is different from the previous attempt but no closer to the correct explanation. An example of the latter type is shown in Figure 1, step 3: although not wrong, it was not necessary to add the term "adjacent".

In assessing this kind of local progress, we obviously need not be concerned with the first attempt at explaining any given step (185 attempts). Further, we disregard steps where the student's input was identical to the previous attempt. Such repetitions occurred rather frequently (namely, 218 times), but we suspect that they are mostly unintentional, due to imperfections in the system's user interface that have been fixed meanwhile. Therefore, we focus on the remaining 388 explanation attempts. For each, we computed whether or not it constitutes progress based on the explanation categories assigned by a human rater, namely, the first author. We found that of the 388 attempts, 44% constituted progress over previous attempts, 26% were classified under the same set of categories as the previous attempt (even though the explanation was different), 17% represented lateral movement, and 14% constituted regression. Thus, the percentage of attempts in which students made progress was close to the 50% threshold presented above, although ideally, it would be somewhat higher than it was.

Some changes have been made within the current dialogue management framework that are likely to improve the progress rate. First, due to its policy of selecting randomly when a student explanation classifies under multiple explanation categories, the tutor did not always base its feedback on the most appropriate category. For example, in step 2 in Figure 1, it would have been better if the tutor had selected category COMPLEMENTARY-ANGLES-90. We have meanwhile changed the tutor's selection criterion so that the tutor selects the explanation category that is closest to a complete and correct explanation. Second, the tutor feedback can likely be improved by associating multiple levels of (increasingly specific) feedback messages with each explanation category. This enables the system to provide more specific feedback when the student's explanation classifies under the same explanation categories as the previous attempt and may help reduce the amount of stagnation in the dialogues. This technique

has since been implemented. Third, some parts of the explanation hierarchy are not yet fine-grained enough to detect certain improvements to explanations. We have added 18 categories already (for a total of 167), based on the analysis of the current data set and will likely add more. Finally, it is likely to be useful if the tutor pointed out to the student whether an explanation attempt implies progress or not. The criterion for progress used here will be a good stating point. While these improvements are very likely to raise the progress rate, we cannot rule out that for some of the more difficult-to-state geometry theorems, such as angle addition, we will need to model more elaborate dialogue strategies that do not easily fit within the current framework.

5 Evaluation of the System's NLU Performance

We also evaluated the system's NLU component, to get an idea of how well it works and because the question of how well knowledge-based natural language understanding works in analyzing mathematical explanations by novices is an interesting research question in its own right. We focused on the accuracy with which the system's NLU component is able to classify student explanations with respect to its set of explanation categories. There is inherent ambiguity in this classification task: for some explanations, it is difficult even for human raters to determine what the correct category or categories are. Therefore, rather than compare the labels assigned by the system against a "correct" set of labels, we ask to what extent the agreement between system and human raters approaches that between human raters. This design was used also in earlier studies, for example a study to evaluate an automated method for essay grading (Foltz, Laham, & Landauer, 1999).

From the set of explanations collected during the pilot study, we removed those explanations that are identical to the previous attempt. As mentioned, we strongly suspect that these repetitions were unintentional, caused by small flaws in the user interface, which meanwhile have been fixed. Therefore, these repetitions should not influence the evaluation of NLU performance. Three human raters (two authors and a research assistant) labeled the remaining set of 573 examples, assigning one or more labels to each explanation. Each rater went through the data twice, in an attempt to achieve maximum accuracy. After the first round, we selected 24 "difficult" examples, explanations that all raters had labeled differently. The raters discussed these examples extensively, in order to calibrate their labeling approaches. We then removed the 24 examples from the data and all raters made a second pass through the data, independently revising their labels. The sets of labels assigned by the human raters and by the system were then processed automatically, in an attempt further to reduce labeling errors, inconsistencies, and irrelevant differences between label sets. These changes preserved the intention of the raters and would not have affected the system's responses. Out of the 167 labels that could be assigned (the categories in the explanation hierarchy plus some extras for explanations that were references to geometry rules), 91 were actually used, combined in 218 different sets of labels.

Table 1. Average pair-wise inter-rater agreement between human raters and average pair-wise agreement between the system and each human rater.

	κ	Actual Agreement	Chance Agreement
Set equality			
Avg Human-Human	0.77	0.77	0.033
Avg System-Human	0.60	0.61	0.030
Overlap			
Avg Human-Human	0.81	0.81	0.043
Avg System-Human	0.65	0.66	0.039
Weighted overlap			
Avg Human-Human	0.88	0.91	0.26
Avg System-Human	0.75	0.81	0.25

To compute the inter-rater agreement, we use the κ statistic (Cohen, 1960), as is customary in the field of computational linguistics and other fields (Carletta, 1996). The κ statistic provides a measure of how much agreement there is between raters beyond the agreement expected to occur by chance alone. We computed κ in three different ways: First we computed κ based on "set equality" – two raters were considered to agree only if they assigned the exact same set of labels to an explanation. However, this measure seems unduly harsh when there are small differences between label sets. Therefore, we also computed two versions of "weighted κ" (Cohen, 1968), a version of the κ statistic that takes into account the degree of difference between labels (or in our case, label sets). So, second, we computed a weighted κ based on "overlap" – meaning that the degree of disagreement was computed as the ratio of the number of unshared labels versus the total number of labels. Third, we computed a weighted κ based on "weighted overlap" – to take into account a (somewhat rough) measure of semantic similarity between the individual labels, measured as the distance in the explanation hierarchy between the labels. In the discussion that follows, we interpret the set equality measure as a lower bound on the agreement and focus mostly on the other two measures.

For each of the three agreement measures, we computed the average of the κ for each pair of human raters, as well as the average of the κ between the system and each human rater. As can be seen in Table 1, the average human-human κ was good. The set equality gives a lower bound of .77. According to the (more appropriate) overlap and weighted overlap measures, the human-human agreement is .81 and .88, respectively. The system-human κs were reasonable but somewhat lower than the corresponding human-human κs. The system-human κ according to the overlap measure is .65, according to the weighted overlap measure, it was .75. Thus, while the comparison of human-human κ and human-system κ indicates that the system's classification accuracy was quite good, there seems to be some room for improvement.

In an attempt to find ways to improve the NLU component, we examined cases where there was high agreement among the human raters (i.e., at least 2 out of the 3 human raters were in full agreement, according to the set equality measure), but where

the system's classification did not agree with the majority of human raters. There were 170 such cases. A detailed examination of those cases revealed about 32 different causes for the system's failure, ranging from difficult to very minor. The most difficult problems deal with insufficient flexibility in the face of ungrammatical language and cases where the system's semantics model was not developed enough to deal with the complexity of the meaning of the student's explanations. The system needs better repair capabilities to deal with ungrammatical sentences such as "the measures of the two angles in the linear pair are add up to 180 degree." Also, the system needs a better semantic representation of coordinated structures, to handle for example the sentence "adjacent supplementary angles form a straight line and are a linear pair."Further, a number of problems of medium difficulty need to be addressed, dealing with quantifiers, relative clauses, multi-clause sentences, and the semantics of certain predicate constructs (e.g., "isosceles triangles have two sides equal"). Finally, there are a number of small flaws with various components of the system that can easily be fixed. A number of problems have been fixed already. While it will take a considerable amount of time to address all problems, overall the evaluation results suggest that knowledge-based NLU with logic-based semantics is able to perform the detailed analysis of student explanations necessary to provide helpful feedback.

6 Discussion and Conclusion

We are developing the Geometry Explanation Tutor in order to evaluate the hypothesis that natural language self-explanation can be tutored effectively by an intelligent tutoring system and leads to improved learning, as compared to alternative ways of supporting self-explanation. We conducted a pilot study to get an idea of the effectiveness of the current system and of the techniques chosen to implement it. This pilot study constituted the first time that the system was used in a school by a group of students from the target population and thus represents a realistic test for the Geometry Explanation Tutor. It does not yet represent a full test, given the limited time and scope; it covered about half the geometry theorems and definitions of the relevant curriculum unit. Also, the students probably were somewhat better prepared and of somewhat higher ability than the students in the target population.

We found evidence that the student-system dialogues are beginning to work well. The logs of student-system dialogues showed evidence that students were learning to explain geometry theorems. On the other hand, the data also revealed some room for improvement. We would like to see a higher completion rate for the dialogues. Also, the number of attempts within dialogues on which students made progress was decent but could be higher. We have described a number of measures that we took in order to improve the system's feedback, which we expect will lead to better progress rates.

Further, there was evidence that the system's knowledge-based NLU component is reaching a reasonably good level of performance in classifying student explanations. We found that the system's classification of student explanations was quite reasonable: it was not too far behind the agreement between human raters, no mean feat

given that we are classifying with respect to a fine-grained set of categories. All in all, the results are encouraging but also indicate room for improvement.

What does the evaluation say about the particular techniques we have chosen for dialogue management and natural language understanding? For some geometry theorems, such as vertical angles, linear pair, and supplementary angles, it seems quite clear that the current dialogue management framework is adequate. It is still an open question however, whether this kind of architecture is going to help students to explain such difficult-to-state theorems as angle addition. At this point, we need to leave open the possibility that we will need more elaborate dialog strategies. Knowledge-based natural language understanding with logic-based semantics seems to be able to deal with challenging input such as students' mathematical explanations.

The broader goal of our project is to get students to learn with greater understanding, as compared to other (simpler) forms of tutored self-explanation. We are currently involved in a larger evaluation study, involving two schools, three teachers, and four classes, in which we compare the current system to a version where students explained their steps by making reference to a rule in a glossary.

References

Aleven, V. (Ed.). (2001). Papers of the AIED-2001 Workshop on Tutorial Dialogue Systems (pp. 59-70). Available via http://www.hcrc.ed.ac.uk/aied2001/workshops.html.

Aleven, V. & Koedinger, K. R. (in press). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science,* 26(2).

Aleven V., Popescu, O., & Koedinger, K. R. (2001). Towards Tutorial Dialog to Support Self-Explanation: Adding Natural Language Understanding to a Cognitive Tutor. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 246-255). Amsterdam, IOS Press.

Aleven, V. & Koedinger, K. R. (2000). The Need for Tutorial Dialog to Support Self-Explanation. In C. P. Rose & R. Freedman (Eds.), *Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium* (pp. 65-73). Technical Report FS-00-01. Menlo Park, CA: AAAI Press.

Allen, J. (1995). *Natural Language Understanding* (2nd Ed.). Redwood City, CA: Cummings.

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.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in Self-Explanation and Self-Regulation Strategies: Investigating the Effects of Knowledge Acquisition Activities on Problem Solving. *Cognition and Instruction,* 13, 221-252.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics,* 22(2):249-254.

Chi, M. T. H. (2000). Self-Explaining Expository Texts: The Dual Processes of Generating Inferences and Repairing Mental Models. In R. Glaser (Ed.), *Advances in Instructional Psychology,* (pp. 161-237). Mahwah, NJ: Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* 70, 213-220.

Conati C. & VanLehn K. (2000). Toward Computer-Based Support of Meta-Cognitive Skills: a Computational Framework to Coach Self-Explanation. *International Journal of Artificial Intelligence in Education,* 11, 398-415.

Evens, M. W., Brandle, S., Chang, R.C., Freedman, R., Glass, M., Lee, Y. H., Shim L.S., Woo, C. W., Zhang, Y., Zhou, Y., Michael, J.A. & Rovick, A. A. (2001). CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue. In *Twelfth Midwest AI and Cognitive Science Conference, MAICS 2001* (pp. 16-23).

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99.*

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine,* 22(4), 39-51.

Heffernan, N. T. & Koedinger, K. R. (2000). Intelligent Tutoring Systems are Missing the Tutor: Building a More Strategic Dialog-Based Tutor. In C. P. Rose & R. Freedman (Eds.), *Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium* (pp. 14-19). Menlo Park, CA: AAAI Press.

Koedinger, K. R., Corbett, A. T., Ritter, S., & Shapiro, L. (2000). *Carnegie Learning's Cognitive Tutor™: Summary Research Results.* White paper. Available from Carnegie Learning Inc., 1200 Penn Avenue, Suite 150, Pittsburgh, PA 15222, E-mail: info@carnegielearning.com, Web: http://www.carnegielearning.com.

MacGregor, R, (1991). The Evolving Technology of Classification-Based Knowledge Representation Systems. In J. Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge.* San Mateo, CA: Morgan Kaufmann.

Popescu, O., & Koedinger, K. R. (2000). Towards Understanding Geometry Explanations In *Proceedings of the AAAI 2000 Fall Symposium, Building Dialog Systems for Tutorial Applications* (pp.80-86). Menlo Park, CA: AAAI Press.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from Worked-Out Examples: the Effects of Example Variability and Elicited Self-Explanations. *Contemporary Educational Psychology, 23,* 90-108.

Renkl, A. (in press). Learning from Worked-Out Examples: Instructional Explanations Supplement Self-Explanations. *Learning and Instruction.*

Rosé, C. P. & R. Freedman, (Eds.). (2000). *Building Dialogue Systems for Tutorial Applications. Papers from the 2000 AAAI Fall Symposium.* Menlo Park, CA: AAAI Press.

Rosé, C. P. & Lavie, A. (1999). LCFlex: An Efficient Robust Left-Corner Parser. User's Guide, Carnegie Mellon University.

Wahlster, W. (2001). Robust Translation of Spontaneous Speech: A Multi-Engine Approach. Invited Paper, *IJCAI-01, Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 1484-1493). San Francisco: Morgan Kaufmann.