The Role of Why Questions in Effective Human Tutoring

C. P. Rosé, D. Bhembe, S. Siler, R. Srivastava, K. VanLehn
Learning Research and Development Center, University of Pittsburgh, Pittsburgh PA, 15260

Abstract.
It is undoubtedly true that one prominent component of effective human tutoring is collaborative dialogue between student and tutor [7, 12]. Nevertheless, many important questions remain to be answered about which features of human tutorial dialogue make it effective and how the most effective human tutoring strategies can be implemented in a tutorial dialogue system. In this paper we present an analysis of a corpus of human tutoring dialogues where we examine which features correlate significantly with learning gains. In particular we explore the role of `Why` questions and other open ended questions in creating opportunities for student learning as well as the role of explicit negative feedback for wrong answers.

1 Introduction

It is undoubtedly true that one prominent component of effective human tutoring is collaborative dialogue between student and tutor [7, 12]. Because of this, in recent years a great deal of interest in building tutorial dialogue systems has developed, and many of the resulting systems have been evaluated successfully with students [15, 6, 5]. Nevertheless, it is still true that important questions remain to be answered about what it is that makes human tutorial dialogue effective and how these effective human tutoring strategies can be implemented in a tutorial dialogue system.

Previous studies have argued the effectiveness of Socratic and other similar tutoring approaches that encourage students to be actively involved in the conversation and say as much as possible. A previous study [16] revealed a trend for Socratic style tutoring dialogues to be more effective for learning than didactic style ones. The rationale behind this result is that students learn more effectively when they are given the opportunity to discover knowledge for themselves [1, 11, 13]. Collins and Stevens (1982) report that the best teachers tend to use a Socratic tutoring style. Recent research on student self-explanations supports the view that when students explain their thinking out loud it enhances their learning [3, 4, 14]. In support of this, in this paper we present an analysis that shows a significant correlation between average student turn length and learning. This leaves open the question of how best to encourage self-explanation from students. In a recent study [8], a computer system was able to elicit effective self-explanations from students as well as human tutors did without understanding the explanations given by students. However, the system used in that study was used solely for the purpose of eliciting self-explanations from students and not offering any feedback. However, in a typical dialogue situation with a human tutor, the tutor offers the student feedback on what the student says. Our hypothesis is that in a dialogue situation where the tutor offers

feedback to the student, it is important for the tutor to understand the student's contributions in order to provide specific feedback for the student as well as to avoid confusing the student by offering inappropriate feedback. The results of our investigation presented in this paper support this hypothesis.

2 WHY Human Tutoring Corpus Collection

The WHY Human Tutoring Corpus is a collection of human tutoring dialogues collected via a Web interface. The corpus features dialogues between students and a human tutor as he coaches them through the process of constructing essays in response to qualitative physics questions such as "Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain."

2.1 Motivation and Background

Recent studies of human tutoring suggest that a productive activity for teaching is to have students explain physical systems qualitatively [2]. We are building a system called WHY2 [18], the purpose of which is to coach students as they explain physics systems in natural language in response to short essay questions such as the one mentioned above. The ultimate goal of WHY2 is to coach students through the process of constructing explanations that are complete and do not contain any misconceptions. As part of our development effort, we collected the WHY Human Tutoring Corpus in order to observe the behavior of an expert human tutor in performing the task that our system is designed to perform.

We first designed a set of 10 essay questions to implement in the WHY2 system. The human tutor in our corpus collection effort uses these same 10 problems in his interaction with the students. As part of the selection process, two physics professors and a computer science professor worked together to outline a set of expectations (i.e., correct propositions that the tutors expected students to include in their essays) and potential misconceptions for each question.

2.2 Experimental Setup

Our expert tutor was instructed to cover the expectations for each problem, and to watch for a set of specific misconceptions associated with the problem. He was trained to avoid lecturing the student and knew that transcripts of their tutoring would be analyzed. He was instructed to end the discussion of each problem by showing an ideal essay to the student. An example dialogue from our corpus is displayed in Figure 1.

For each problem, the student was first presented with an essay question. After reading the essay question, the student typed in an initial essay into the essay input window, which is in the lower right hand corner of the interface. The tutor then analyzed the student's essay. Based on his analysis, he then engaged the student in a dialogue addressing gaps in the student's understanding of the underlying physics and coaching the student on ways to improve the essay. At key points in the dialogue, the tutor asked the student to revise the essay. When the tutor was satisfied with the student's essay, he showed the student the ideal essay. When the student indicated being finished reading the ideal essay, the tutor then moved on to the next problem.

PROBLEM: Suppose that you released 3 identical balls of clay in a vacuum at exactly the same instant. They would all hit the ground at the same instant. Now you stick two of the balls together, forming one ball that is twice as heavy as the remaining, untouched clay ball. Both balls are released in a vacuum at exactly the same instant. Which ball hits the ground first?

ESSAY: Both balls will hit the ground at the same time. The balls are in free fall (only gravitational forces). The ratio of the masses and weight are equal.

. . . excerpt from 2 minutes into a typed dialogue . . .

Tutor1: You have correctly said the balls are in free fall. What do you conclude from this fact?

Student1: There is no air resistance and the balls' accelerations are constant in regards to one another

Tutor2: Right, so the conclusion is that they both have the same acceleration, the free fall acceleration. Now you have to show that time taken by them to reach the ground will be the same. How do you do that?

Student2: F (net forces) / mass = f/m because F = m*a therefore ratio should equal one another

Tutor3: But that does not tell you anything about the time taken. What physical quantities will determine the time taken. Will the initial velocity of the balls matter?

Student3: yes, assuming starting at rest? sorry, kind of at a loss at the moment

Tutor4: The initial velocity will matter, and here they both have the same initial velocity, zero. What else, will the height through wich they fall matter?

Student4: not so much as long as they are both dropped from the same height

Tutor5: Height of fall will determine the time of fall, and here the height of fall is the same for both the balls. Now on the basis of these facts can you conclude about the time taken by them to reach the ground?

Student5: since initial velocities are equal and the height of falls are equal, then the time taken is equal

Tutor6: How about acceleration, if they had different accelerations, even if they fall through the same height and have same initial velocity , will they reach the ground at the same time?

Student6: no...

Tutor7: Right...

Figure 1: **Excerpt from WHY Human Tutoring Corpus**

2.3 Experimental Design

The subjects were University of Pittsburgh students who had never taken a college level physics course before. The experiment took altogether no more than 9 hours per student, which was split up between 1 and 3 sessions at the end of the Fall 2002 semester. 17 students began the study, but only 12 subjects have completed the study so far. Some subject were not able to complete the study at the end of the Fall semester due to scheduling conflicts between the students and the human tutor. We are currently continuing to collect and analyze data.

Two tests were developed as pre/post tests: versions A and B, which were isomorphic to one another. That is, the problems on test A and B differed only in the identities of the objects (e.g., cars vs. trucks) and other surface features that should not affect the reasoning required to solve them. Each version of the test (A and B) consisted of 4 essay questions and 40 multiple choice questions. Essays questions were written that addressed the same expectations as were targeted in the training problems. These were similar in style to the ones used for the 10 training problems. Thus, they each covered multiple expectations. Each multiple choice question was written to address a single expectation covered in the training problems.

Because we used subjects who had never taken college level physics before, after taking the pretest, subjects then read through a 9-page document summarizing some physics background material that would be helpful to them as they worked through the 10 training problems. This material was extracted from chapters 2-6 of [9]. The subjects then worked through the 10 training problems with the human tutor. After they completed the problems, they were given the post-test. Some students were not able to complete all 10 problems before they reached the end of their participate time. Thus, they took the post-test after only working through a subset of the training problems.

2.4 Results

The results of the essay portion of the pre and post tests are still being analyzed, thus we based our analysis here solely on the multiple-choice portion of the test. We present scores as fractions of 1, indicating the percentage of points earned by each student. For the twelve students who completed both the pre-test and the post-test, the average pre-test score was 0.47 with standard deviation 0.10. Average post-test score was .66 with standard deviation .14. We present a detailed analysis of the transcripts for the first 7 students who have completed the study.

Because of ample prior research supporting the view that when students explain their thinking out loud it enhances their learning [3, 4, 14], we ran linear regression tests over the transcripts in order to investigate correlations between Average number of words per student turn and post-test score, number of essay words and post-test score, and total number of student dialogue words and post-test score. For average turn length, the overall average student turn length was 13.88 words with standard deviation 5.55 words. A linear regression with independent variable average student turn length for each student and dependent variable post-test score, with pre-test score regressed out, yielded a significant result ($R = 0.911; p < .01$). Using number of words used to formulate the student's essay for each student over their entire experience with the human tutor as the independent variable, we did not find a significant result (p=.653). Since we found a significant correlation between average student turn length and learning, we tested whether the number of words uttered by the student over

the course of the whole dialogue correlated with learning. The linear regression with average number of dialogue words for each student correlated with learning did not yield a significant result (p=.244). The results of this preliminary analysis left us with the question of why the average turn length would play a significant role, but the total number of words uttered either in in the dialogue or in the essays did not.

One possible hypothesis is that students who uttered longer student turns had more prior knowledge, and thus more to say in response to each question. However, when we ran a linear regression to investigate the correlation between pretest score and average turn length, we did not find a significant result (p=.898). A second hypothesis is that the student's average turn length is influenced by the strategy of the tutor, and that this strategy for eliciting longer student turns leads to more learning on the part of the student.

In order to investigate this hypothesis, we first sorted pairs of Tutor/Student turns according to the length of the student's turn in order to examine which types of tutor questions elicited the longest student answers. We found that 20% of the corpus consisted of student turns that were 20 words long or greater. Half of these were student turns that were 30 words or longer. 60% of the corpus consisted of turns that were 10 words long or less. Thus, there was a great deal of variation in length on tutor turns, and the great majority of them were fairly short, i.e., well below the overall average student turn length. We informally examined the top 10% of student turns to see which types of tutor questions were used to elicit them. 57% of these were open ended questions, 35% of which were `Why` questions.

Because of this informal observation, in order to explore the contribution of the tutor's strategy to the strong correlation between average student turn length and learning, we first coded our transcripts for tutor question type within the discussion portions of the transcripts. This includes all but the student essays themselves except for tutor turns immediately preceding essay revision, since these did not have any corresponding student response. Altogether the discussion portion of the transcripts for the seven students includes 505 student-tutor exchanges. Each exchange is composed of a tutor question, potentially preceded by a short explanation, a student answer, and normally some feedback from the tutor.

Questions were assigned one of six classifications:

Understanding Questions that asked whether the student understood the preceding explanation, i.e., "Are you ready to move on?", or "Do you have any questions?"

Why Questions that ask for a causal explanation that can naturally be preceded by "because" or "the reason is". This includes questions such as "How do you conclude...?" or "What factors determine...?"

Closed This includes yes/no questions, questions that ask for a single canonical fact such as the statement of a rule or definition of a technical term, or questions that can be answered by rephrasing the question as a statement and replacing the wh-expression with a phrase not preceding by "because". For example, "Which ball will land first?"

Reask Questions that rephrase an earlier Closed question.

Open All other questions. Examples of open questions include "What would you conclude from this?" "What did you mean by the opposition reaction?"

Meta Meta discussion pertaining to essay revision, viewing the ideal essay, and moving between problems.

The distinction between Why/Open questions and Closed questions in this analysis is similar in spirit to that described in [10], which proved to be a distinction that could be reliably coded. The most frequent question class was `Closed`. There were 290 closed questions, making up 57% of the questions. The next largest class of questions were `Meta` questions. These were 113 of these questions, making up 22% of the questions. The next largest class was `Why` questions. There were 55 of these, making up 11% of the questions. After that came other `Open` questions. There were 37 of these, making up 7% of the questions. Thus, altogether there were 92 open ended questions, 60% of which were `Why` questions. Thus, `Why` questions were the most frequent open ended questions in our corpus. The two smallest classes were `Understanding`, of which there were 32, and `Reask`, of which there were 15.

We looked at the relationship between question type and length of student response. While it was true that answers to open ended questions were sometimes very short, and answers to closed class questions were many times lengthy, on average answers to open ended questions were longer than answers to closed class questions. The average student turn length in response to open ended questions was 21.98, with a standard deviation of 16.93, whereas the mean for other question types was 12.08, with a standard deviation of 12.88. ($t(503)=6.266$, $p<.01$). Because student answers to open ended questions were longer on average than other question types, we checked to see whether the percentage of `Why` questions overall for each student reliably predicted either post-test score or average student turn length. However, both a linear regression between percentage of `Why` questions and average student turn length, as well as between percentage of `Why` questions and post test score, with pretest score regressed out, came out non-significant. Because `Why` questions are relatively infrequent overall, although they are the most frequent open ended questions, it is not surprising that concentration of `Why` questions alone did not account for average student turn length or learning.

We then coded tutor responses for types of negative feedback after an incorrect student answer. We coded for three types of explicit negative feedback, which were not mutually exclusive. For illustrative purposes, let's assume that the student has incorrectly claimed that gravitational force acts in the horizontal direction. Feedback marked as `Pointing` indicated that the tutor explicitly pointed out something wrong in the student's response, either by stating that it was wrong, directly questioning it, as in "Is gravity a horizontal force?", or by stating the opposite of what the student said, as in "Gravitational force does not act in the horizontal direction." `Negative` was assigned to a tutor response if the tutor began by saying, "No", or "That is not right." We did not count tutor turns preceded by "Well" in this class. `Right` was assigned to a tutor response if the tutor corrected the student's incorrect statement in his turn, as in "Gravitational force acts in the *vertical* direction." `Ignore` was assigned to tutor responses to wrong answers that did not contain any of the above types of explicit negative feedback. These turns included cases where the tutor simply rephrased his original question or tried a different question altogether. Often, the different question seemed to be meant to direct the student toward the correct answer to the original question. However, it was not explicitly indicated as such. We consider the `Ignore` responses to be implicit negative feedback. Evidence from occasional student comments in the corpus indicate that students viewed them this way as well.

We then tested whether students were more likely to receive some form of negative feedback after an open ended or closed class question. For this analysis we did not consider `Meta` questions since these questions do not specifically target physics content. For each student we computed the percentage of time that a student received negative feedback after an

open ended question and the percentage of time the student received negative feedback after a closed class question. On average, students received negative feedback 57% of the time after an open ended question, with a standard deviation of 14%. In contrast, students received negative feedback after a closed class question 43% of the time, with a standard deviation of 8%. The difference was marginal ($t(5)=1.98$, $p=.1$). Note that we did not include numbers from one student who received only 1 open ended question during his entire experience with the tutor. Thus, students were more likely to answer open ended questions incorrectly. One possible interpretation is that students are less likely to say something wrong when they give short answers, since answers to closed class questions were shorter on average.

We then looked closer at patterns of negative feedback. One observation we made was that it was very rare for the tutor to say, "No" to students. Only 10% of the instances of negative feedback were preceeded by, "No" or "That is wrong.". The tutor most frequently gave some form of explicit negative feedback. Implicit negative feedback was given only in 22% of the total number of cases of negative feedback in the corpus. We computed the frequency of implicit negative feedback for each student separately. We found that students received implicit negative feedback on average in 17% of the cases where they received negative feedback, with a standard deviation of 9%. Since we informally observed students offering lengthy justifications when explicit negative feedback was offered, we tested for a correlation between percentage of wrong answers that received explicit negative feedback and average student turn length. We found a moderately reliable correlation ($R=.5105$, $p=.07$). Thus, while we do not conclude from this that explicit negative feedback is the primary factor determining average student turn length, we suspect that it is one important factor.

3 Conclusions and Current Directions

In this paper we presented an analysis of a corpus of human tutoring dialogues where we examine patterns of six different question types and four different forms of negative feedback for wrong answers. One interpretation of this data is as follows: Both open ended questions and explicit negative feedback together encourage students to say more. When students say more, they are more likely to say something wrong. In response to wrong answers, the tutor offers instruction. Most of the time, the tutor offers explicit instruction when students say something wrong. Thus, saying something wrong is useful since it creates an opportunity for the student to receive instruction. Explicit negative feedback was more effective than implicit negative feedback in eliciting longer explanations from students, perhaps because explicit negative feedback is a more face threatening behavior.

The analysis reported in this paper was conducted by a single coder over transcripts from only 7 students who have completed the WHY Human tutoring study. We plan to do a more extensive analysis involving more transcripts and multiple coders, in order to verify the reliability of the coding and to further explore the role of open ended questions and explicit negative feedback in learning.

4 Acknowledgments

This research was supported by the Office of Naval Research, Cognitive Science Division under grant number N00014-0-1-0600 and by NSF grant number 9720359 to CIRCLE, Center for Interdisciplinary Research on Constructive Learning Environments at the University of

Pittsburgh and Carnegie Mellon University.

References

[1] A. L. Brown and M. J. Kane. Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20:493–523, 1988.

[2] M. Chi, N. de Leeuw, M. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 1981.

[3] M. T. H. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182, 1989.

[4] M. T. H. Chi, N. de Leeuw, M. H. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, 1994.

[5] M. Evans and J. Michael. *One-on-One Tutoring by Humans and Machines*. Lawrence Earlbaum and Associates, in-press.

[6] A. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, and the Tutoring Research Group. Autotutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1(1):35–51, 1999.

[7] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. Collaborative dialogue patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology*, 9:495–522, 1995.

[8] R. G. M. Hausmann and M. T. H. Chi. Can a computer interface support self-explanation. *The International Journal of Cognitive Technology*, 7(1):4–13, 2002.

[9] P. G. Hewitt. *Conceptual Physics*. Adison Wesley, 1987.

[10] Pamela Jordan and Stephanie Siler. Student initiative and questioning strategies in computer-mediated human tutoring dialogues. 2002.

[11] M. C. Lovett. Learning by problem solving versus by examples: The benefits of generating and receiving information. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. NJ:Erlbaum, 1992.

[12] Douglas C. Merrill, Brian J. Reiser, and S. Landes. Human tutoring: Pedagogical strategies and learning outcomes, 1992. Paper presented at the annual meeting of the American Educational Research Association.

[13] M. Pressley, E. Wood, V. E. Woloshyn, V. Martin, A. King, and D. Menke. Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27:91–109, 1992.

[14] A. Renkl. Worked-out examples: Instructional explanations support learning by self-explanations, submitted.

[15] C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. Interactive conceptual tutoring in atlas-andes. In *Proceedings of Artificial Intelligence in Education*, pages 256–266, 2001.

[16] C. P. Rosé, J. D. Moore, K. VanLehn, and D. Allbritton. A comparative evaluation of socratic versus didactic tutoring. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 869–874, 2000.

[17] Albert L. Stevens, Allan Collins, and Sarah E. Goldin. Misconceptions in student's understanding. *International Journal of Man-Machine Studies*, 11:145–156, 1979.

[18] K. VanLehn, P. Jordan, C. P. Rosé, and The Natural Language Tutoring Group. The architecture of why2-atlas: a coach for qualitative physics essay writing. Proceedings of the Intelligent Tutoring Systems Conference, 2002.