Evidence of Misunderstandings in Tutorial Dialogue and their Impact on Learning [1]

Pamela JORDAN [a,2], Diane LITMAN [a,b], Michael LIPSCHULTZ [b] and Joanna DRUMMOND [b]

[a] *Learning Research and Development Center, University of Pittsburgh*
[b] *Department of Computer Science, University of Pittsburgh*

Abstract. We explore the frequency and impact of misunderstandings in an existing corpus of tutorial dialogues in which a student appears to get an interpretation that is not in line with what the system developers intended. We found that this type of error is frequent, regardless of whether student input is typed or spoken, and that it does not respond well to general misconception repair strategies. Further we found that it is feasible to detect misunderstandings and suggest alternative strategies for repairing them that we intend to test in the future.

Keywords. misunderstandings, tutorial dialogue, dialogue systems

Introduction

When a dialogue system and user interact, both must interpret an input signal, whether spoken or typed, to recognize words, phrases, sentences and finally the intended meaning given the context of what has been previously communicated. The hearer of the signal must then react appropriately in response. Interpretation errors can happen at any stage of the recognition process. With speech input, research typically focuses on detecting and repairing speech recognition errors that occur in the early stages of interpretation. With typed input, which is predominant still for tutorial dialogue systems (TDS), the focus is more toward the final cognitive/pragmatic stages of the interpretation process [14]. In this paper we explore whether there are errors that happen between these two extremes in tutorial dialogue that could negatively impact learning.

We will use the term *misconception* (e.g. [10,5,1]) for knowledge errors or deficits users may have about the domain and the term *misunderstanding* to refer to a hearer getting an interpretation that is not in line with what the speaker intended [13,4]. The hearer may or may not suspect he has misunderstood. If he does, he will typically attempt a repair in response. If the hearer does not, then his unexpected response can be a signal that a previous contribution was misunderstood.

[1] This work was funded in part by NSF awards 0631930 and 0428472
[2] Email: pjordan@pitt.edu

system:	Now, if the car's magnitude of velocity is 4 times greater than the old one, how much greater would its kinetic energy be?
student:	The new KEO would be 15,984,720 J *(categories: 4 times,* **attempted***, no attempt*
system:	I disagree with you. If the car's magnitude of velocity increases 4 times then the new magnitude is equal to v*4 where v represents the car's old magnitude of velocity.
system:	Using the formula for kinetic energy, you can see that the new kinetic energy will be 1/2*m*(4*v)^2 which can be rewritten as: 16*(1/2*m*v^2)= 16*KE.

Table 1. Example of an unexpectedly wrong response that could be due to misunderstanding the question as requesting a quantitative rather than qualitative solution

Misunderstandings can arise from divergences in the participants' views of what has been established in the conversational record [6]. These divergences can arise because of e.g. mis-remembering, failing to remember, or simple performance errors in problem solving or in language, such as mis-speaking or mis-hearing. For spoken dialogue systems, mis-hearing is reflected by the word recognition error rate (WER). High WERs do negatively correlate with user satisfaction for tutorial applications but do not correlate with learning [9,11]. In this paper we provide evidence that regardless of whether input is typed or is spoken with human-quality speech recognition available, other types of potential misunderstandings are frequent and can have a negative impact on learning. Mishearing is probably less of an issue in tutoring since the system will still take an action that is most likely useful for the student, assuming the student can correctly interpret the system's follow-up [12].

While analyzing both spoken and typed tutorial dialogues between students and TDSs, we found a significant number of unexpected student responses that appear not to be due to misconceptions. When we annotated these unexpected responses by type, we found that students whose responses were either unexpectedly wrong (i.e. a wrong response not anticipated by the developers), as in Table 1, or non-attempts to respond did not benefit as much from a repair intended for a misconception as those students whose wrong responses were expected (i.e. responses that are more clearly misconceptions).

First we will describe the Cordillera corpus of typed dialogues that we analyzed. Next we will present the evidence that suggests that misunderstandings are happening frequently, can negatively impact learning and do not respond well to misconception repairs. Next we briefly describe the ITSPOKE spoken dialogue corpus that we analyzed and again show that possible misunderstandings are frequent. We then describe how we used this corpus to explore the feasibility of detecting misunderstandings. Finally, we conclude with a discussion of possible repair strategies that we intend to test in the future.

1. The Cordillera Tutorial Dialogue System and Corpus

The Cordillera tutorial dialogue system tutors students in both quantitative and qualitative physics in the work-energy domain. Cordillera was developed to provide a testbed for exploring questions about the relationship between tutorial dialogue strategies and student learning [15] and was implemented using the TuTalk TDS toolkit [7]. TuTalk supports the authoring and delivery of dialogues in which a tutor tries to elicit the main line of reasoning from a student by a series of coherent questions. This style of dialogue was inspired by CIRCSIM-Tutor's directed lines of reasoning [2].

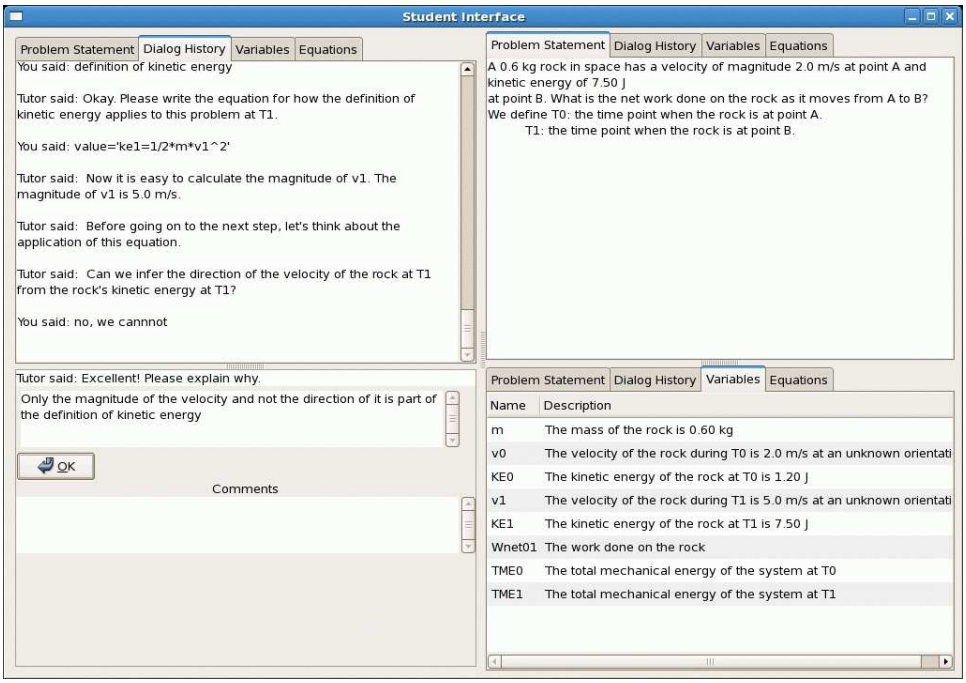

Figure 1. The student interface

Domain experts authored dialogues for seven problems that guided students through the problems by either hinting at or telling the student the next problem solving step that needed to be completed. Hints were usually in the form of short answer questions. Figure 1 illustrates a sample student dialogue with Cordillera. The upper right pane of the figure shows the problem that the student is attempting to solve. The top left pane shows a portion of the dialogue history, and illustrates a few questions and student responses, as well as a number of system informs; the pending tutor question is shown in the input pane at the bottom. Finally, the variables in the bottom right pane were defined either by the student using a form interface or provided by the tutor.

In addition to guiding the student through problem solving, Cordillera also tried to help increase the student's conceptual understanding by asking for justifications and exploring how changes in the problem statement effect the solution. While the system's requests in the problem solving discussions were short answer questions (see the first and last tutor turn in the top left pane), requests for justifications were usually in the form of deep answer questions (see the current tutor turn in the bottom left pane).

To reduce the confounds of imperfect natural language (NL) understanding on their investigations of how different dialogue strategies impact student learning, the fully automated (but error-prone) TuTalk NL understanding module was replaced with a human language understanding *wizard*. The wizard's interface mirrors that of the student, except that the bottom input pane is replaced by the student's response and a set of check-boxes for classifying the student's response.

The Cordillera corpus is a collection of 448 typed dialogues between students and Cordillera, and associated corpus annotations. The corpus consists of a database of computer-logged turns and system state, including how the wizard classified the student

system1:	What is the equation for the work done by the man on the crate?
student1:	? [**ic**] [**parallel_work**]
system2:	Recall that the general equation for work involves force and displacement. In this case the respective variable names are W_Fa, Fa, and d. So, just piece them together: What is the equation for the work done by the man on the crate?
student2:	Wa=Fa*d [**c**] [**parallel_work**]
	Dialogue goes on

Table 2. Sample coded dialog excerpt.

turns (e.g., as an *attempted* response) and which physics knowledge components (KCs) were covered, as will be explained below. A dialogue excerpt with an unexpectedly wrong response from the corpus is shown in Table 1.

64 students (all physics novices) were recruited from Pittsburgh university campuses. Each student read a short introductory text on work-energy and then took a 37 item test (pre-test), where each question was associated with one or more KCs. Next, the students completed a demonstration problem that introduced them to Cordillera's interface and then worked with Cordillera on the seven training problems. The training took about 7-10 hours spread across multiple days. When students completed the training problems, they took the same test again (post-test). The pre/post-test was designed to be able to compute learning gains for each KC.

As the system interacted with students it logged which KCs were covered by the system's turns or the system-student turn pairs. For example, for the student-system dialogue in Table 2, given input from the language understanding wizard, the system recorded that for the first tutor/student turn pair the student was unable to evoke the KC about work being the magnitude of the force times the displacement, while in the last tutor/student pair the student exhibited the correct KC. Thus, the first student turn was logged as [**ic**] to indicate it is incorrect and [**parallel_work**] for the KC and the last student turn as [**c**] to indicate it is correct and [**parallel_work**] again for the KC. Students' learning gains were significant for a majority of the KCs, as were their composite learning gains.

In the next section, we show evidence both that *unexpected* responses are high frequency and that some types of them impact students ability to learn.

2. Frequency and Impact of Misunderstandings

Our overall goal for analyzing the Cordillera corpus was to gain an understanding of the *unexpected* responses. We found that it wasn't that a set of tutor turns were particularly problematic; many students would respond as expected for those same tutor turns. We had various insights after looking at the dialogues; one being that some of the students knew the answer but had misunderstood the question given their past performance and the content of their answer. And in some cases the student would report that they had misunderstood the question after being told the expected answer.

In the Cordillera corpus only two *unexpected* categories were available for the wizards to select; *attempted* and *no attempt*. For responses wizards classified as *no attempt*, e.g. "I don't know", the system gave no explicit feedback on correctness. For responses wizards classified as *attempted*, negative feedback was given since the response did not answer the intended question. In both cases, the subsequent repair was the same, the

student was told the correct expected response. An example of a response classified as *attempted* is shown in Table 1. The italicized list following the answer represents the categories available to the wizard and the one in bold represents the one selected by the wizard.

We found that 21% of all NL responses from students were classified as *attempted* and 12% as *no attempt*. There were significant weak negative correlations between post-test scores (after removing the effects of pre-test scores) and the percentage of student's responses that were classified as *no attempt* (R=-.30,p=.017) or *attempted* (R=-.32,p=.011) and a significant moderate positive correlation for the percentage of student's expected responses (R=.47,p=0). The expected responses are step specific and are either correct responses or non-correct responses that warrant a specific follow-up.

During this initial examination of the corpus we identified the following alternative categories for covering *unexpected* responses; *no attempt*, *wrong*, *vague* and *overly specific* which are related to typical response classes described in tutorial dialogue literature [2,8]. We had two physics knowledgeable students annotate one third of the Cordillera corpus' unexpected responses with these alternative categories with a Kappa of .65. After one student annotated the remaining unexpected responses for the seven most frequently discussed KCs, we found that 60% were annotated as *wrong*, 21% as *vague*, 11% as overly-specific, 7% as *no attempt* and for 1% the annotator disagreed with the wizard and considered them *correct*.

We expected that some of the *wrong*, *overly specific* and *no attempt* responses, the latter of which are similar to a time-out in spoken dialogue systems, could be due to misunderstandings. Intuition suggests that *vague* and *overly specific* responses could indicate the student is close to having understood both the question and the KC. The remaining *unexpected* responses suggest that the student may not yet have understood. This lack of understanding could be of the KC or of the system's turn and request. The first is an issue for tutoring and human learning and is domain dependent while the latter is an issue for dialogue systems and communication in general. If it is the first case then telling the student the correct expected response should be an adequate repair. But if it is the second case then this same repair may not be adequate. That is, the student may continue to have difficulty understanding the way the system communicates (i.e. it will never make its intentions explicit).

To test these ideas we looked at the learning curves for the seven KCs combined. We subdivided students into groups according to how their first opportunity for a KC was classified and used the wizard's original classifications for expected responses. For each group we generated five separate learning curves for the categories of *correct*, *overly specific*, *vague*, *wrong* and *no attempt* by averaging error rates for the seven KCs. After a first opportunity we plotted the percentage of those same students who were not correct on that KC on the second opportunity, then the third and so on. We then generated the learning curves to fit this data as shown in Figure 2. All of the fits were significant. When we look at the curves we see that the error rate drops for all the groups of students as they have repeated opportunities to discuss a KC. But the error rate at the right-hand side of the graph is higher for the subgroups whose first opportunities were either *wrong* (top line) or *no attempt* (next line down) than for the other groups of students. So the repair of telling the student the correct expected response did not help as much for these two groups of students. Perhaps some of the students were helped because they lacked knowledge while others may have lacked an understanding of the communicative intentions.

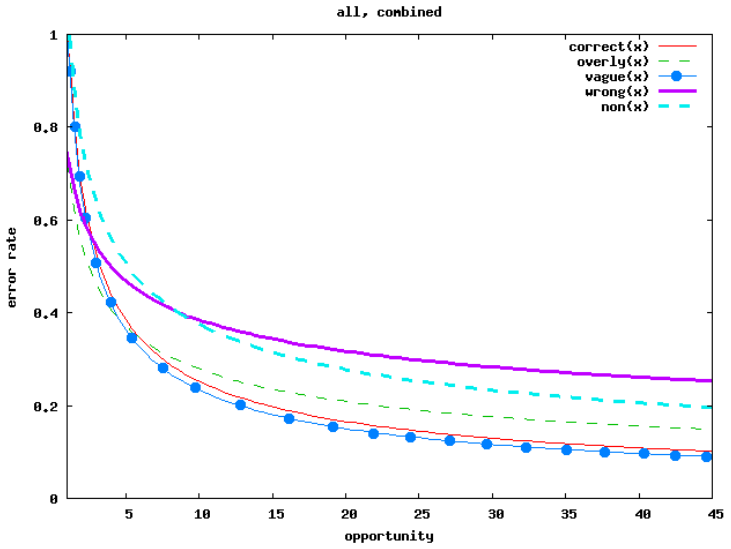

Figure 2. Cordillera error rates based on classification of first opportunities

We next compared the group of students whose first response for two KCs had been classified as wrong by the wizard (i.e. these responses were expectedly wrong and have an error specific follow-up) with those students whose first response was classified as unexpectedly wrong. We see, as shown in Figure 3, that there is not a significant fit for the learning curve for this KC when an initial response was unexpectedly wrong (the top line on the right-hand side). But for those students whose first response was expectedly wrong there is a significant fit and the learning curve indicates that the students who received the error specific remediations did have a decrease in their error rate (the bottom line on the right-hand side). For the second KC the comparison was similar.[3] For the expectedly wrong responses the user presumably understood the question and had a misconception while for the unexpectedly wrong responses a repair that assumed a misconception was not helpful.

3. Feasibility of Detecting Misunderstandings

Our second annotator also marked the *unexpected* responses in an ITSPOKE system corpus of spoken tutorial dialogues. For this corpus collection a wizard again did the language understanding for the dialogue system and the corpus is more fully described in [3]. The domain was about a different area of physics than the Cordillera corpus and all the problem solving was qualitative rather than a mixture of qualitative and quantitative. 19% of all student responses in this corpus were classified as *unexpected* by the wizard. We found that 79% of these *unexpected* responses were annotated as wrong. Again these errors could not be due to the system mis-hearing the student because a wizard

[3]The other five KCs did not have any expectedly wrong classifications to compare against.

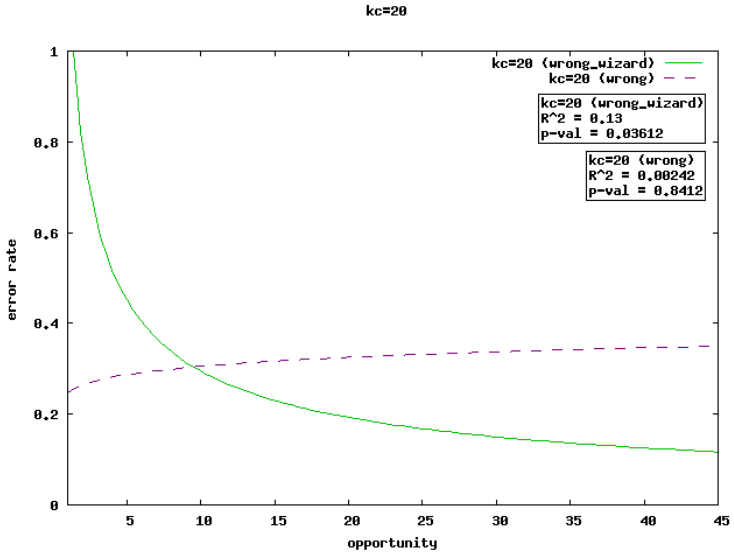

Figure 3. Cordillera error rates for one KC for expectedly vs unexpectedly wrong responses

did speech and language understanding. So these errors had to be due to either lack of domain knowledge or a misunderstanding[4].

We next used the corpus with the new annotations for unexpected responses to check the feasibility of detecting the unexpected categories and to explore which automatically available features, including spoken language features such as pitch and duration, would be useful for automatic classification. In addition the ITSPOKE corpus had already been manually annotated for uncertainty [3] and we wanted to explore whether uncertainty could help distinguish between categories of *unexpected* responses.

Because the set of *unexpected* responses was heavily skewed toward *wrong* responses (79%) we used re-sampling to balance the corpus. We found that when using spoken language features and word level features, such as the text of the tutor's prior turn and the word edit distance between the student response and the expected correct response, the classifier achieved an accuracy of 50% which was better than the majority class baseline accuracy of 37.6%. With word level features alone the classifier achieved an accuracy of 43%. With spoken language features alone the accuracy during training was high but was not better than the baseline during testing. The strength of the word level features alone suggest that it is feasible to train a classifier that is sensitive to misunderstandings that may need a different type of repair but we will need to explore additional features to improve the accuracy.

4. Conclusions

In this paper we showed evidence that suggests it is important to have strategies that can generally detect and repair misunderstandings that arise from more than just mis-hearing

[4]This corpus has not yet been annotated for KCs so we were unable to do a learning curve analysis.

what is said since these errors 1) can dominate for complex domains (e.g. tutoring vs. booking a flight) and 2) can be detrimental to the user's task success.

After a preliminary examination of some of the unexpectedly wrong responses, we hypothesize that potential misunderstandings could arise from at least three sources; 1) not recognizing implicit intentions about problem solving and reasoning strategies 2) not utilizing all of the relevant previously mentioned information needed to get the system's intended interpretation and 3) not recognizing how the system's request connected with the immediately previous dialogue. For each of these hypothesized sources we have identified some general repair strategies that can be applied across domains and applications, such as making intentions explicit, that we intend to test in the future for their impact on learning.

References

[1] R. J. Calistri-Yeh. Utilizing user models to handle ambiguity and misconceptions in robust plan recognition. *User Modelling and User-Adapted Interaction*, 1(4):289–322, 1991.

[2] M. Evens and J. Michael. *One-on-One Tutoring by Humans and Computers*. Lawrence Erlbaum Associates, Inc., 2006.

[3] K. Forbes-Riley, D. Litman, and M. Rotaru. Responding to student uncertainty during computer tutoring: An experimental evaluation. In *Proceedings of Intelligent Tutoring Systems Conference (ITS 2008)*, pages 60–69, 2008.

[4] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton. Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15(3–4):213–229, 1994.

[5] H. Horacek and M. Wolska. Interpreting semi-formal utterances in dialogs about mathematical proofs. In F. Meziane and E. Métais, editors, *Natural Language Processing and Information Systems*, volume 3136 of *LNCS*, pages 26–38. Springer, 2004.

[6] P. Jordan and R. Thomason. Refining the categories of miscommunication. In *Proceedings of AAAI Workshop on Detecting, Repairing, and Preventing Human-machine Miscommunication*, CA, 1996. AAAI Press.

[7] P. W. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C.P. Rosé. Tools for authoring a dialogue agent that participates in learning studies. In *Proceedings of AIED 2007*, pages 43–50, 2007.

[8] H.C. Lane. *Natural Language Tutoring and the Novice Programmer*. PhD thesis, University of Pittsburgh, Department of Computer Science, 2004.

[9] D. J. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16:145–170, 2006.

[10] M. Makatchev, P. Jordan, and K. VanLehn. Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems. *Journal of Automated Reasoning, Special issue on Automated Reasoning and Theorem Proving in Education*, 32:187–226, 2004.

[11] H. Pon-Barry, K. Schultz, E.O. Bratt, B. Clark, and S. Peters. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16(2):171–194, 2006.

[12] M. Rotaru and D. Litman. Dependencies between student state and speech recognition problems in spoken tutoring dialogues. In *Proceedings of Coling/ACL*, Sydney, Australia, 2006.

[13] G. Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45:325–341, 2005.

[14] K. VanLehn. The behavior of tutoring systems. *International Journal of Artificial Intelligence and Education*, 16, 2006.

[15] K. VanLehn, P. Jordan, and D. Litman. Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proceedings of SLaTE Workshop on Speech and Language Technology in Education*, 2007.