Dialogue-Learning Correlations in Spoken Dialogue Tutoring

Kate Forbes-Riley [a] Diane Litman [a] Alison Huettner [a] and Arthur Ward [a]

[a] *University of Pittsburgh, Learning Research and Development Center, 3939 O'Hara Street, Pittsburgh, PA, 15260, USA.*

Abstract. We examine correlations between dialogue characteristics and learning in two corpora of spoken tutoring dialogues: a human-human corpus and a human-computer corpus, both of which have been manually annotated with *dialogue acts* relative to the tutoring domain. The results from our human-computer corpus show that the presence of student utterances that display reasoning, as well as the presence of reasoning questions asked by the computer tutor, both positively correlate with learning. The results from our human-human corpus show that the introduction of a new concept into the dialogue by students positively correlates with learning, but student attempts at deeper reasoning do not, and the human tutor's attempts to direct the dialogue negatively correlate with learning.

1. Introduction

Research in tutorial dialogue systems is founded on the belief that a one-on-one natural language conversation with a tutor provides students with an environment that exhibits characteristics associated with learning. However, it is not yet well understood exactly how specific student and tutor dialogue behaviors correlate with learning, and whether such correlations generalize across different types of tutoring situations.

In the computational tutoring community, understanding such correlations has become of increasing interest, in order to put system building on a more empirical basis; this is because when it comes time to actually implement a tutorial dialogue system, many design choices must be made that will likely influence the style of the dialogue, which in turn may influence a student's ability to learn from the system. One area of interest has been the use of shallow measures to investigate the hypothesis that increased student language production correlates with learning; shallow measures have the advantage of being automatically computable, and are thus easy to incorporate into an online adaptive system. Studies of typed (primarily human-human) dialogue tutoring corpora, for example, have shown that longer student turns, and higher percentages of student words and student turns, all positively correlate with learning [1,2,3].

Unfortunately, when in prior work we applied similar measures to other types of tutoring dialogues - namely *spoken* dialogues, and human-*computer* dialogues (typed and spoken) - we found that although our students learned, most correlations between learning and shallow dialogue measures did not generalize to our data [4]. Furthermore, even when some shallow correlations did generalize (as in our typed human-human data), we felt that further analysis was still needed to better understand our results. For example, one might hypothesize that longer student turns are a good estimate of how much a student explains, but a deeper coding of the data would be needed to test this hypothesis.

In fact, the notion of a "dialogue act" [5,6,7], which attempts to codify the underlying intent behind a student or tutor utterance, has been used in recent studies of both implemented [8] and simulated [9] computer tutors. For example, the correlation studies of [8] suggest that student learning is positively correlated with the use of tutor dialogue acts requiring students to provide the majority of an answer, and negatively correlated with the use of tutor acts where the tutor primarily provides the answer.[1]

In this paper, we take a similar approach, and analyze correlations between learning and dialogue acts. However, we examine learning correlations with both *tutor* as well as *student* dialogue acts. In addition, we examine and contrast our findings across two types of spoken dialogue corpora: one with a *human* tutor, and the other with a *computer* tutor. The results in our human-computer corpus show that the presence of student utterances that display reasoning, as well as the presence of reasoning questions asked by the computer tutor, both positively correlate with learning. The results from our human-human corpus are more complex, mirroring the greater complexity of human-human interaction: the introduction of a new concept into the dialogue by students positively correlates with learning, but student attempts at deeper reasoning do not, and the human tutor's attempts to direct the dialogue can negatively correlate with student learning.

2. Dialogue Data and Coding Schemes

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) [11] is a *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system [12]. Our data consists of two corpora of spoken tutoring dialogues, one with the ITSPOKE computer tutor, and the other with a human tutor performing the same task as ITSPOKE. Both corpora were collected during a prior study [4], using the same experimental procedure: university students 1) took a pretest measuring their physics knowledge, 2) read a small document of background material, 3) used a web and voice interface to work through a set of up to 10 training problems (dialogues) with the tutor, and 4) took a posttest similar to the pretest.[2] In each training problem, students first typed an essay answering a qualitative physics problem; the tutor then engaged the student in spoken dialogue to correct misconceptions and elicit more complete explanations. Annotated (see below) examples from our two corpora are shown in Figures 1 and 2 (punctuation added for clarity).[3]

For our current study, each *tutor turn* and each *student turn* in these two corpora was manually annotated for tutoring-specific dialogue acts.[4] Our tagset of "Student and Tutor Dialogue Acts" is shown and briefly defined in Figure 3. This tagset was developed based on pilot annotation studies using similar tagsets previously applied in other tutorial dialogue projects [13,5,6,7]. As shown, "Tutor and Student Question Acts" label the

[1] Correlations between similar codings of dialogue data have also been studied in collaborative learning research. For example, [10] shows that students who more often indicated that they needed help by asking specific questions learned more than those who asked fewer specific questions (R= 0.48, $p < .01$).

[2] In the human-computer corpus, students worked through 5 problems, and took the pretest after the reading.

[3] The *human-computer* corpus contains 100 dialogues (20 students), averaging 22 student turns and 29 tutor turns per dialogue. The *human-human* corpus contains 128 dialogues (14 students), averaging 47 student turns and 43 tutor turns per dialogue.

[4] While one annotator labeled the entire corpus, an agreement study on a subset of the corpus gave 0.67 Kappa and 0.63 Kappa between two annotators on 334 tutor turns and 442 student turns, respectively.

type of question that is asked, in terms of content and the expectation that the content presupposes with respect to the type of answer required. This Act is most common to the tutor; as detailed below, there are no student questions in our human-computer corpus, and they are infrequent in our human-human corpus. “Tutor Feedback Acts” essentially label the “correctness” of the student’s prior turn, in terms of explicit positive or negative tutor responses. “Tutor State Acts” serve to summarize or clarify the current state of the student’s argument, based on the prior student turn(s). “Student Answer Acts” label the type of answer that a student gives, in terms of the quantity and quality of the content and the extent of reasoning that the content requires. Finally, the “NonSubstantive Act” (NS) tag was used to label turns that did not contribute to the physics discussion (e.g., “Are you ready to begin?”).

. . . spoken dialogue excerpt at 2.6 minutes into session. . .

STUDENT$_{5}$: Gravitational pull, downward. *[NOVS]*

ITSPOKE$_{6}$: Almost. The forces acting on the pumpkin are both vertical forces. The force exerted by the person is in the upward direction. The person also exerts a force on the pumpkin. One of the forces on the pumpkin is gravity. We have established that there are two forces acting on the pumpkin, gravity in the downward direction and the force exerted by the person in the upward direction. So what is the direction of the NET force? *[BOT, RS, RC, SAQ]*

STUDENT$_{7}$: Upward. *[NOVS]*

ITSPOKE$_{8}$: Yeah. Does the pumpkin accelerate during the toss? If yes, please specify the direction of the acceleration. *[POS, SAQ, RD]*

STUDENT$_{9}$: Yes, upward. *[SHAL]*

ITSPOKE$_{10}$: Well... To analyze the pumpkin’s acceleration we will use Newton’s second law. What is the definition of this law? *[HNT, LAQ]*

Figure 1. Annotated Human-Computer Dialogue Excerpt

. . . spoken dialogue excerpt at 16.6 minutes into session. . .

TUTOR$_{101}$: Yeah, that’s precisely the point. Yes, we are all learning. Ok so uh now uh you apply the same push for the same amount of time for on both the containers then what would you compare to distinguish between them? *[POS, RC, SAQ]*

STUDENT$_{102}$: I would be comparing their rate of velocity. *[NOVS]*

TUTOR$_{103}$: Not rate. You will be comparing their velocity. You see, rate will imply that something is changing which there is no change. Velocity is constant so you will surely compare their velocities. Which one will be faster? *[HNT, RS, SAQ]*

STUDENT$_{104}$: The feathers. *[NOVS]*

TUTOR$_{105}$: The feathers- why? *[RS, DAQ]*

STUDENT$_{106}$: Because there’s less matter. *[DEEP]*

Figure 2. Annotated Human-Human Dialogue Excerpt

As Figures 1-2 illustrate, most tutor turns are labeled with multiple Tutor Acts, while most student turns are labeled with a single Student Act. Applying the Dialogue Act coding scheme to our human-computer corpus yielded 2293 Student Acts on 2291 student turns and 6879 Tutor Acts on 2964 tutor turns. Applying the coding scheme to our human-human corpus yielded 5969 Student Acts on 5879 student turns and 7861 Tutor Acts on 4868 tutor turns.

- **Tutor and Student Question Acts**
  - Short Answer Question (**SAQ**): concerns basic quantitative relationships.
  - Long Answer Question (**LAQ**): requires definition/interpretation of concepts.
  - Deep Answer Question (**DAQ**): requires reasoning about causes and/or effects.
- **Tutor Feedback Acts**
  - Positive Feedback (**POS**): overt positive response to prior student turn.
  - Negative Feedback (**NEG**): overt negative response to prior student turn.
- **Tutor State Acts**
  - Restatement (**RS**): repetitions and rewordings of prior student statement.
  - Recap (**RC**): restating student's overall argument or earlier-established points.
  - Request/Directive (**RD**): directions summarizing expectations about student's argument.
  - Bottom Out (**BOT**): complete answer supplied after student answer is incorrect or incomplete.
  - Hint (**HNT**): partial answer supplied after student answer is incorrect or incomplete.
  - Expansion (**EX**): novel details about student answer supplied without first being queried.
- **Student Answer Acts**
  - Deep Answer (**DEEP**): consists of at least two concepts linked together through reasoning.
  - Novel/Single Answer (**NOVS**): consists of one concept introduced by student into dialogue.
  - Shallow Answer (**SHAL**): consists of one concept previously introduced into dialogue.
  - Assertion (**AS**): used for answers such as "I don't know" or equivalent.
- **Tutor and Student Non-Substantive Acts (NS)**: do not contribute to the physics discussion.

Figure 3. Student and Tutor Dialogue Acts

3. Correlation Analysis Methodology

As discussed in Section 1, although our prior work demonstrated that students learned a significant amount with both our human and computer tutors [4], in our spoken data we were unable to find any correlations between learning and a set of shallow dialogue measures of increased student activity (e.g., longer student turns). Here we revisit the question of what aspects of our spoken dialogues correlate with learning, but replace our previous shallow measures for characterizing dialogue with a set of "deeper" measures derived from the Student and Tutor Dialogue Act annotations described in Section 2.

For each of our two corpora, we first computed for each student, a total, a percentage, and a ratio representing the usage of each Student and Tutor Dialogue Act tag across all of the dialogues with that student. We call these measures our *Dialogue Act Measures*. Each *Tag Total* was computed by counting the number of (student or tutor) turns that contained that tag at least once. Each *Tag Percentage* was computed by dividing the tag's total by the total number of (student or tutor) turns. Finally, each *Tag Ratio* was computed by dividing the tag's total by the total number of (student or tutor) turns that contained a tag of that tag *type*. For example, suppose the dialogue in Figure 1 constituted our entire corpus. Then our Dialogue Act Measures for the Tutor "POS" tag would be: Tag Total = 1, since 1 tutor turn contains the "POS" tag. Tag Percentage = 1/3, since there are 3 tutor turns. Tag Ratio = 1/1, since 1 tutor turn contains a Tutor Feedback Act tag.

Next, for each of the Dialogue Act Measures, we computed a Pearson's correlation between the measure and posttest score. However, because the pretest and posttest scores were significantly correlated in both the human-human (R=.72, p =.008) and human-computer corpora (R=.46, p=.04), we controlled for pretest score by regressing it out of the correlation.[5] In the following Sections (4 and 5), we present and discuss the best results of these correlation analyses, namely those where the correlation with learning was significant ($p \leq .05$) or a trend ($p \leq .1$), after regressing out pretest.

4. Human-Computer Results

Table 1 presents our best results on correlations between Dialogue Act Measures and learning in our human-computer corpus. The first column lists the measure (total (#), percentage (%) or ratio (Rat:) of the Dialogue Act per student). The second and third columns show the mean and standard deviation (across all students), while the last two columns present the Pearson's correlation between posttest and the measure after the correlation with pretest is regressed out. For example, the first row shows that there are 11.90 total Deep Answers over all the dialogues of a student on average, and that there is a statistically significant (p=.04) positive correlation (R = .48) between total Deep Answers and posttest, after the correlation with pretest is regressed out.

Table 1. Dialogue-Learning Correlations: Human-Computer Corpus (20 students)

Dialogue Act Measure	Mean	Std.Dev.	R	p
# Student DEEP	11.90	5.78	.48	.04
# Tutor DAQ	9.59	4.89	.41	.08
% Tutor DAQ	6.27%	2.30%	.45	.05
% Tutor Question Acts	76.89%	3.12%	.57	.01
Rat: Tutor SAQ to Question Acts	.88	.04	-.47	.04
Rat: Tutor DAQ to Question Acts	.08	.03	.42	.07
# Tutor POS	76.10	16.66	.38	.10

As shown, the *type of answer provided by students* relates to how much they learn in our human-computer corpus, as indicated by the positive correlation between student Deep Answers and learning. Note that there are no significant (positive or negative) correlations for student Shallow or Novel/Single Answers, or a student's inability to provide an answer (Assertions), which suggests that the relationship between student answer type and learning requires further analysis.

The *type of questions asked by tutors* also relates to how much students learn in our human-computer corpus. There is a positive correlation between the percent of tutor Deep Answer Questions and learning, and a trend for the number and ratio of tutor Deep Answer Questions to positively correlate with learning. In contrast, there is a negative correlation between the ratio of tutor Short Answer Questions and learning. The *quantity of tutor questions* also relates to student learning, as evidenced by the strong positive correlation between the overall percentage of all tutor Question Acts and learning.

Table 1 also shows a slight trend for tutor Positive Feedback to positively correlate with learning. Other studies have shown positive relationships between encouragement during computer tutoring and student outcomes [14]. Finally, note that none of the tutor

[5] The human-human means for the (multiple-choice) pre- and posttests were 0.42 and 0.72, respectively, and the human-computer means were 0.48 and 0.69, respectively.

State Acts correlated with learning, suggesting that the best way to use such organizational acts is not yet fully understood in our computer tutor.

5. Human-Human Results

Table 2 presents our best results on correlations between Dialogue Act Measures and learning in our human-human corpus, using the same format as Table 1. As shown, the *type of dialogue acts used by students* relates to how much students learn in our human-human corpus too. With respect to student answers, here we find a trend for the number and ratio of student Novel/Single Answers to positively correlate with learning; however, in contrast to our human-computer results, we also find a trend for the number of student Deep Answers to *negatively* correlate with learning. Moreover, unlike in the human-computer corpus, in our human-human corpus students do ask questions. Here we see that a higher ratio of student Short Answer Questions positively correlates with learning, and a higher ratio of student Long Answer Questions negatively correlates with learning.

Table 2. Dialogue-Learning Correlations: Human-Human Corpus (14 students)

Dialogue Act Measure	Mean	Std.Dev.	R	p
# Student NOVS	19.29	7.95	.49	.09
# Student DEEP	68.50	27.99	-.49	.09
Rat: Student NOVS to Answers	.14	.05	.47	.10
Rat: Student SAQ to Question Acts	.91	.12	.56	.05
Rat: Student LAQ to Questions	.03	.08	-.57	.04
# Tutor RD	19.86	10.58	-.71	.01
% Tutor RD	5.65%	.02	-.61	.03
# Tutor RS	79.14	26.83	-.56	.05
# Tutor NEG	14.50	7.60	-.60	.03

Table 2 also shows that the *type of dialogue acts used by the tutor* relates to how much students learn in our human-human corpus. In contrast to the human-computer corpus, in our human tutoring dialogues we only find correlations with non-question tutor Acts (namely State Acts and Negative Feedback), and also find only negative correlations. The correlations between tutor State Acts (RD, RS) and learning show that increased tutor summarization and clarification negatively correlates with student learning. We also see a negative correlation between tutor Negative Feedback and learning.

6. Discussion

Our human-human corpus represents an upper bound for the speech and natural language processing capabilities of our ITSPOKE corpus. As such, cross-corpora differences in how student and tutor dialogue acts relate to student learning can shed light on how system improvements might positively impact learning. We see little overlap in terms of the correlations between tutoring Dialogue Acts and learning across our human-computer and human-computer corpora. In our computer tutoring data, we found that student learning was positively correlated with both the presence of student utterances displaying reasoning, as well as the presence of tutor questions requiring reasoning. These results are similar to previous findings in human-tutoring data, where learning was correlated with both students' construction of knowledge, and tutor behaviors prompting students to construct knowledge [13]. We hypothesize that because Deep Answers involve more student reasoning, they involve more knowledge construction. Note that we previously found no significant correlation between average turn length (# words/turn) or dialogue length (total words) and learning in either our human-computer or human-human corpora [4]; together these results suggest that it is not the quantity but the quality of the students' responses that correlate with learning.

The results from our human-human corpus are more complex. First, there is no longer a straightforward correlation between the depth of reasoning displayed in student answers and learning: while student Novel/Single insights positively correlate with learning, student attempts at even deeper reasoning negatively correlate with learning. While this negative correlation is surprising, inspection of the student turns in the human-human corpus leads us to hypothesize that student Deep Answers might often be incorrect, which itself might negatively correlate with learning, and may also be related to the fact that in the human-human corpus, students speak longer and more freely than in the human-computer corpus. We are currently annotating "correctness", to investigate whether more Deep Answers are "incorrect" or "partially correct" in the human tutoring corpus compared to the computer tutoring corpus, and whether the number of correct answers positively correlates with learning. Similarly, the correlations between tutor Feedback and learning in both corpora might also reflect correctness. Second, while student question-asking is often considered a constructive activity [13], we similarly did not see a straightforward relation between question-asking and learning: while student Short Answer Questions positively correlate with learning, student Long Answer Questions negatively correlate. However, there were only 12 Long Answer Questions in our human-human data, and all displayed clear evidence of student misunderstanding (e.g., containing phrases such as "what do you mean?"). Finally, although we find negative correlations between learning and tutor State Acts (e.g., involving summarization and clarification), attributing any causal relationship would require further research.

Finally, we see some overlap between our results and those of [8], who computed correlations between student learning and tutor dialogue acts in the AutoTutor system. [8] found that students who received more "Hints" (which require the student to provide most of the answer) learned more than those who received more "Assertions" (in which the tutor provides most of the answer). Although our Tutor Act coding is not identical, our "Bottom Out" largely corresponds to their "Assertion"; in our human-human corpus there was a non-significant negative correlation (R=-.00,p=.99), but in our human-computer corpus there was a non-significant positive correlation (R=.08, p=.75), with learning. Our "Hint" is similar to their "Hint"; in our human-computer corpus there was a non-significant positive correlation (R=.26, p=.28), but in our human-human corpus there was a non-significant negative correlation (R=-.38,p=.20), with learning.

7. Conclusions and Current Directions

This paper presented our findings regarding the correlation of student and tutor dialogue acts with learning, in both human-human and human-computer spoken tutoring dialogues. Although we found significant correlations and trends in both corpora, the results for specific dialogue acts differed. This suggests the importance of training systems from appropriate data. The results in our human-computer corpus show that student utterances that display reasoning, as well as tutor questions that ask for student reasoning, both positively correlate with learning. The results in our human-human corpus mirror the greater complexity of human-human interaction: student novel insights positively correlate with learning, but student deeper reasoning is negatively correlated with learning, as are some of the human tutor's attempts to direct the dialogue. As noted above, to gain further insight into our results, we are currently annotating our dialogues for correctness. This will allow us to test our hypothesis that student deep reasoning is more error-prone in the human-human corpus. We are also investigating correlations between learning and *patterns* of dialogue acts, as found in multi-level coding schemes such as [7].

Acknowledgments

We thank Mihai Rotaru and Pam Jordan for their help in improving this paper. This research is supported by ONR (N00014-04-1-0108) and NSF (0325054).

References

[1] M. G. Core, J. D. Moore, and C. Zinn. The role of initiative in tutorial dialogue. In *Proc. European Chap. Assoc. Computational Linguistics*, 2003.

[2] C. P. Rosé, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn. The role of why questions in effective human tutoring. In *Proceedings of Artificial Intelligence in Education*, 2003.

[3] Sandra Katz, David Allbritton, and Johen Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence and Education*, 13, 2003.

[4] D. J. Litman, C. P. Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. Spoken versus typed human and computer dialogue tutoring. In *Proc. Intell. Tutoring Systems*, 2004.

[5] A. Graesser and N. Person. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137, 1994.

[6] A. Graesser, N. Person, and J. Magliano. Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9:495–522, 1995.

[7] R. M. Pilkington. Analysing educational discourse: The DISCOUNT scheme. Computer-Based Learning Unit 99/2, University of Leeds, 1999.

[8] G. Jackson, N. Person, and A. Graesser. Adaptive tutorial dialogue in AutoTutor. In *Proc. Workshop on Dialog-based Intelligent Tutoring Systems at Intelligent Tutoring Systems*, 2004.

[9] M. Wolska, B. Q. Vo, D. Tsovaltzi, I. Kruiff-Korbayová, E. Karagjosova, H Horacek, A. Fiedler, and C. Benzmuller. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *Proc. Language Resources and Evaluation*, 2004.

[10] N. Webb and A. M. Mastergeorge. The development of student helping behavior and learning in small groups. *Cognition and Instruction*, 21(4):361–428, 2003.

[11] D. Litman and S. Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc Human Language Technology: North American Chap. Assoc. Computational Linguistics*, 2004.

[12] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems*, 2002.

[13] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471–533, 2001.

[14] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. In *Proc. Intelligent Tutoring Systems*, 2002.