Natural Language Engineering **12** (2): 161–176. © 2006 Cambridge University Press
doi:10.1017/S1351324906004165 Printed in the United Kingdom

Correlations between dialogue acts and learning in spoken tutoring dialogues

DIANE LITMAN and KATE FORBES-RILEY

Learning Research and Development Center,
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: litman@cs.pitt.edu

(*Received 12 November 2005*)

Abstract

We examine correlations between dialogue behaviors and learning in tutoring, using two corpora of spoken tutoring dialogues: a human-human corpus and a human-computer corpus. To formalize the notion of dialogue behavior, we manually annotate our data using a tagset of student and tutor *dialogue acts* relative to the tutoring domain. A *unigram* analysis of our annotated data shows that student learning correlates both with the tutor's dialogue acts and with the student's dialogue acts. A *bigram* analysis shows that student learning also correlates with joint patterns of tutor and student dialogue acts. In particular, our human-computer results show that the presence of student utterances that display reasoning (whether correct or incorrect), as well as the presence of reasoning questions asked by the computer tutor, both positively correlate with learning. Our human-human results show that student introductions of a new concept into the dialogue positively correlates with learning, but student attempts at deeper reasoning (particularly when incorrect), and the human tutor's attempts to direct the dialogue, both negatively correlate with learning. These results suggest that while the use of dialogue act n-grams is a promising method for examining correlations between dialogue behavior and learning, specific findings can differ in human versus computer tutoring, with the latter better motivating adaptive strategies for implementation.

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 this generalizes across different types of tutoring situations.

In the computational tutoring community, understanding such correlations is of increasing interest, 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 (Core, Moore and Zinn 2003; Rosé, Bhembe, Siler, Srivastava and VanLehn 2003; Katz, Allbritton and Connelly 2003).

Unfortunately, when in prior work we applied similar measures to other types of tutoring dialogues – namely *spoken* dialogues, and human-*computer* dialogues (typed or spoken) – we found that although our students learned, most correlations between learning and shallow dialogue measures did not generalize to our data (Litman, Rosé, Forbes-Riley, VanLehn, Bhembe and Silliman 2004). Furthermore, even when shallow correlations did generalize (as in our typed human-human data), further analysis was needed. 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, some computational studies have started to analyze dialogue-learning correlations with a variety of deeper codings. In particular, the notion of a "dialogue act" (Graesser and Person 1994; Graesser, Person and Magliano 1995; Pilkington 1999; Core and Allen 1997), which attempts to codify the underlying intent behind a student or tutor utterance, has been used in recent studies of both implemented (Jackson, Person and Graesser 2004) and simulated (Wolska, Vo, Tsovalyzi, Kruiff-Korbayová, Karagjosova, Horacek, Fiedler and Benzmuller 2004) computer tutors. Jackson *et al.* (2004) 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.

In this paper we take a similar approach, and analyze dialogue act-learning correlations. However, we examine learning correlations with both *tutor* and *student* dialogue acts, and we analyze learning correlations with both *single occurrences* and *sequences* of these dialogue acts. Our analyses of student dialogue act unigrams, tutor dialogue act unigrams, student-tutor dialogue act bigrams and tutor-student dialogue act bigrams represent our approach to exploring three main hypotheses in the literature regarding why one-on-one dialogue tutoring is so effective: a student-centered hypothesis, a tutor-centered hypothesis, and an interaction-centered hypothesis (Chi, Siler, Jeong, Yamauchi and Hausmann 2001). Our dialogue act analyses lend support for all three hypotheses: in our corpora, student learning is correlated with student dialogue acts, with tutor dialogue acts, and with bigrams containing both types of 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. Given the current limitations of human language technologies, the use of human tutors provides a benchmark for estimating the performance of an "ideal" computer system with respect to such technologies. Our human-computer corpus results show that the presence of student answers that display reasoning (whether correct or incorrect), as well as reasoning questions asked by the computer tutor, both positively correlate

. . . spoken dialogue excerpt at 2.6 minutes into session. . .
ITSPOKE$_6$: 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? *[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? *[HN, LAQ]*

Fig. 1. Annotated human-computer dialogue excerpt.

. . . spoken dialogue excerpt at 16.6 minutes into session. . .
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? *[HN, RS, SAQ]*
STUDENT$_{104}$: The feathers. *[NOVS]*
TUTOR$_{105}$: The feathers- why? *[RS, DAQ]*
STUDENT$_{106}$: Because there's less matter. *[DEEP]*
TUTOR$_{107}$: No no- because they accelerated more. *[NEG, BO]*

Fig. 2. Annotated human-human dialogue excerpt.

with learning, both alone and in the context of larger dialogue patterns. Student answers that reiterate concepts also correlate with increased learning, but only in the context of particular tutor acts. Our human-human corpus results mirror the greater complexity of human-human interaction: student answers that introduce new concepts positively correlate with learning, but student attempts at deeper reasoning (particularly when incorrect) and some human tutor attempts to direct the dialogue both negatively correlate with learning, alone and within larger dialogue patterns. Student questions also negatively correlate with learning in the context of particular tutor dialogue acts. In sum, our results suggest that while the use of dialogue act n-grams is a promising method for exploring correlations between dialogue behavior and learning, specific findings can differ in human versus computer tutoring, with the latter better motivating adaptive strategies for implementation.

2 Dialogue data and coding schemes

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman and Silliman 2004) is a *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system (VanLehn *et al.* 2002). Our data consists of two corpora of spoken tutoring dialogues, one with the ITSPOKE tutor, and the other with a human tutor performing the same task as ITSPOKE.

The corpora were collected during two prior studies (Litman *et al.* 2004) using a similar experimental procedure to enable qualitative cross-corpora comparisons:

university students (1) took a pretest measuring their physics knowledge, (2) used a web and voice interface to work through a set of training problems (dialogues) with the tutor, and (3) took a posttest similar to the pretest. Students also read a small document of background material to prepare them for the training problems. In the human-computer study, students took the pretest after the reading and then worked through five problems with the computer tutor. In the human-human study, students took the pretest before the reading and then worked 7–10 problems with the human tutor. Each dialogue begins after the student types an essay answering a qualitative physics problem; the tutor then engages the student in spoken dialogue to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. Annotated (see below) examples from our corpora are shown in Figures 1 and 2 (punctuation added for clarity). The human-computer corpus contains 100 dialogues from 20 students, averaging 22 student turns and 29 tutor turns per dialogue.[1] The human-human corpus contains 128 dialogues from 14 students, averaging 47 student turns and 43 tutor turns per dialogue.

For our current study, each tutor and student utterance in these two corpora was manually annotated for tutoring-specific dialogue acts. Our tagset of "Student and Tutor Dialogue Acts" is shown and briefly defined in Figure 3. As shown, "Tutor and Student Question Acts" label the 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" label feedback based on the presence of *lexical items* in the tutor turn. Although these tags often coincide with the correctness of a student turn, they can also convey encouragement, or relate to the discourse level or to the student's earlier essay. "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 the student gives, in terms of the quantity and quality of the content and the extent of reasoning that the content requires. "NonSubstantive Acts" label turns that did not contribute to the physics discussion (e.g., backchanneling or social coordinations); these occur mainly in the human-human corpus. This tagset was developed based on pilot studies using similar tagsets from other tutorial dialogue projects (Chi *et al.* 2001; Graesser and Person 1994; Graesser *et al.* 1995; Pilkington 1999). Note that since tutoring dialogues have a number of tutoring-specific dialogue acts (e.g., hinting), tutorial dialogue research typically uses tutoring-specific dialogue act tagsets rather

[1] ITSPOKE's word error rate in this corpus was 31.2%, but natural language understanding based on speech recognition (i.e., recognition of semantic concepts instead of actual words) is the same as based on perfect transcription 92.4% of the time. An internal evaluation of this semantic analysis component in an early version of the Why2-Atlas system (with typed input, and thus "perfect" transcription) yielded 97% accuracy. The accuracy of recognizing semantic concepts is more relevant for dialogue evaluation, as it does not penalize for word errors that are unimportant to overall utterance interpretation.

- **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)**: positive feedback *lexical item* is present in the turn
  - — Negative Feedback **(NEG)**: negative feedback *lexical item* is present in the 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 expected student argument
  - — Bottom Out **(BO)**: full answer given if student answer is incorrect/incomplete
  - — Hint **(HN)**: partial answer given if student answer is incorrect/incomplete
  - — Expansion **(EX)**: novel details about answer given without being queried
- **Student Answer Acts**
  - — Deep Answer **(DEEP)**: at least two concepts linked together through reasoning
  - — Novel/Single Answer **(NOVS)**: one concept student introduces into dialogue
  - — Shallow Answer **(SHAL)**: one concept previously introduced into dialogue
  - — Can't Answer **(CA)**: 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

Fig. 3. Student and tutor dialogue acts.

Table 1. *Student (S) and tutor (T) dialogue act totals in human-computer (C) and human-human (H) corpora*

	S:	SAQ	LAQ	DAQ	DEEP	NOVS	SHAL	CA	NS
	C	0	0	0	238	323	1528	94	111
	H	125	12	7	979	272	797	168	3640

T:	SAQ	LAQ	DAQ	POS	NEG	RS	RC	RD	BO	HN	EX	NS
C	2011	131	190	1522	41	492	515	208	516	400	635	251
H	1205	167	586	902	203	1273	305	298	336	865	1061	1279

than more domain-independent schemes such as DAMSL (Core and Allen 1997) (although DAMSL is still somewhat biased towards task-oriented dialogue).[2]

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 student turns yielded 2294 Student Acts on 2291 student

[2] (Rickel *et al.* 2001) presents a first step towards bridging this gap, by integrating an initial set of tutoring-specific acts into a more general collaborative discourse framework. (Wolska *et al.* 2004) are extending DAMSL to address the needs of tutoring.

Table 2. *Confusion matrix for tutor act inter-annotation*

	DAQ	LAQ	SAQ	BO	EX	HN	RD	RC	RS	NEG	POS	NS
DAQ	25	1	14	0	0	2	2	0	0	0	0	4
LAQ	1	4	1	0	0	0	0	0	0	0	0	1
SAQ	3	9	60	0	0	2	1	0	1	0	0	7
BO	0	0	1	12	2	0	1	1	3	0	0	4
EX	0	0	0	0	18	3	6	11	0	0	0	20
HN	0	1	0	2	8	9	11	7	0	16	0	17
RD	0	0	0	0	0	0	5	0	0	0	0	1
RC	0	0	1	0	2	0	0	14	1	0	0	4
RS	0	0	0	3	1	1	2	13	35	1	1	21
NEG	0	0	0	0	0	0	0	0	0	9	0	4
POS	0	0	0	0	0	0	0	0	0	1	34	18
NS	0	0	1	1	0	0	10	0	2	3	0	68

turns in our human-computer corpus and 6000 Student Acts on 5879 student turns in our human-human corpus, distributed as shown in the top of Table 1. Applying the Dialogue Act coding scheme to our tutor turns yielded 6912 Tutor Acts on 2964 tutor turns in our human-computer corpus and 8480 Tutor Acts on 4868 tutor turns in our human-human corpus, as shown in the bottom of Table 1.

While one annotator labeled both corpora in their entirety, an agreement study was performed between two annotators who separately annotated 8 human-human dialogues containing 459 Student Acts and 548 Tutor Acts. The confusion matrices in Tables 2-3 summarize inter-annotator agreement across these Tutor Acts (Kappa = 0.48, "Moderate")[3], and Student Acts (Kappa = .68, "Substantial"). The agreement for Tutor Acts, which with 12 categories is the harder annotation task, improves (Kappa = .63, "Substantial") if we collapse the 12 categories into Question Acts, State Acts, Feedback Acts, and NonSubstantive Acts.

3 Dialogue-learning correlation analysis methodology

In this paper we investigate what types of dialogue behaviors correlate with student learning, replacing our previously investigated shallow measures for characterizing dialogue (e.g. turn length) (Litman *et al.* 2004) with a set of unigram and bigram measures derived from the "deeper" Student and Tutor Dialogue Act annotations described in Section 2.[4] First, to explore the student and tutor-centered hypotheses discussed in Section 1, we computed a set of *Unigram Dialogue Act Measures*. In particular, for each student in our two corpora, we computed a total, a percentage, and a ratio representing the use of each Student and Tutor Dialogue Act tag across all the dialogues with that student. 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

[3] Kappa value interpretation is somewhat controversial and varies depending on the application field. Landis and Koch (1977) and others use this agreement standard: 0.21–0.40, "Fair"; 0.41–0.60, "Moderate"; 0.61–0.80, "Substantial"; 0.81–1.00, "Almost Perfect".

[4] Our unigram results in this paper were first published in Forbes-Riley *et al.* (2005).

Table 3. *Confusion matrix for student act inter-annotation*

	DEEP	NOVS	SHAL	CA	DAQ	LAQ	SAQ	NS
DEEP	38	3	14	0	0	0	1	5
NOVS	1	9	2	0	0	0	0	1
SHAL	6	16	31	1	0	0	0	11
CA	0	0	0	14	0	0	0	2
DAQ	0	0	0	0	0	0	0	0
LAQ	0	0	0	0	0	0	0	0
SAQ	0	0	0	0	0	0	1	3
NS	0	2	6	1	0	0	2	289

or tutor) turns. 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*. Suppose the dialogue in Figure 1 constituted our entire corpus. Then our Unigram 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.

Second, to explore the interaction-centered hypothesis noted in Section 1, we also computed a set of *Bigram Dialogue Act Measures*. These measures quantify the presence of two specific types of larger dialogue patterns in our corpora, namely *sequences* of two dialogue acts containing *both* a Student and a Tutor Dialogue Act:

- **Student Act**$_n$ **- Tutor Act**$_{n+1}$: all bigrams constructed by pairing each Student Dialogue Act in turn n with each Tutor Dialogue Act in turn $n+1$
- **Tutor Act**$_n$ **- Student Act**$_{n+1}$: all bigrams constructed by pairing each Tutor Dialogue Act in turn n with each Student Dialogue Act in turn $n+1$

In particular, for each student in each corpora, we computed the *total* number of occurrences of each instantiation of the two different Dialogue Act Bigram types across all of the dialogues with that student.[5] Suppose the dialogue in Figure 1 constituted our entire human-computer corpus. Consider the **Student Act**$_n$ **- Tutor Act**$_{n+1}$ turn sequences, of which there are 2 in Figure 1. In each of the tutor turns in these sequences, there are multiple Dialogue Acts, thus there are multiple bigrams computed for each sequence. For example, for **STUDENT**$_7$ **- ITSPOKE**$_8$, there are 3 bigrams computed (bracketed hereafter for clarity): **[Student NOVS - Tutor POS]**, **[Student NOVS - Tutor SAQ]**, and **[Student NOVS - Tutor RD]**. Similarly, there are 2 **Tutor Act**$_n$ **- Student Act**$_{n+1}$ turn sequences in Figure 1, and again there are multiple bigrams computed for each sequence. As this example indicates, we create separate bigrams from each Dialogue Act tag in a turn, rather than treating the entire sequence of Dialogue Acts per turn as a single element in the bigram. This is because there are no limits on tag combinations per turn, thus treating each tagged turn as a unique bigram element would yield a data sparsity problem for our analysis

[5] Due to space limitations we did not compute percentages or ratios for the bigrams as we did for the unigrams, since the denominator could be either element of the bigram.

of correlations between bigrams and learning. As a result, we decided to consider the presence or absence of each Dialogue Act tag separately when constructing our Bigram Dialogue Act Measures.

Finally, for each of the Unigram and Bigram 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=.65, p =.01) and the human-computer corpus (R = 0.46, p = 0.04), we controlled for pretest score by regressing it out of the correlation.[6]

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 ($p \leq 0.05$) or a trend ($p \leq 0.1$), after regressing out pretest.

4 Human-computer results

Table 4 presents our best results on correlations between Unigram 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. The groupings of rows represent the student and tutor dialogue acts results, respectively. 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 ($p = 0.04$) positive correlation ($R = 0.48$) between total Deep Answers and posttest, after the correlation with pretest is regressed out.

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 (Can't Answers), suggesting 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; 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 4 also shows a trend for tutor Positive Feedback to positively correlate with learning. Finally, none of the tutor State Acts correlated with learning; this suggests

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

Table 4. *Unigram-learning correlations: human-computer corpus (20 students)*

Unigram Dialogue Act Measure	Mean	Std.Dev.	R	p
# Student DEEP	11.90	5.78	0.48	0.04
# Tutor DAQ	9.59	4.89	0.41	0.08
% Tutor DAQ	6.27	2.30	0.45	0.05
% Tutor Question Acts	76.89	3.12	0.57	0.01
Rat: Tutor SAQ to Question Acts	0.88	0.04	−0.47	0.04
Rat: Tutor DAQ to Question Acts	0.08	0.03	0.42	0.07
# Tutor POS	76.10	16.66	0.38	0.10

Table 5. *Bigram-learning correlations: human-computer corpus (20 students)*

Bigram Dialogue Act Measure	Mean	Std.Dev.	R	p
# [Student DEEP - Tutor SAQ]	9.10	4.42	0.49	0.03
# [Student DEEP - Tutor DAQ]	1.80	1.91	0.41	0.08
# [Student DEEP - Tutor RS]	4.75	2.00	0.41	0.08
# [Student DEEP - Tutor BO]	3.50	3.07	0.42	0.07
# [Student DEEP - Tutor EX]	4.20	2.59	0.46	0.05
# [Student SHAL - Tutor DAQ]	4.95	2.58	0.39	0.10
# [Student SHAL - Tutor LAQ]	4.00	1.78	0.40	0.09
# [Student SHAL - Tutor POS]	49.65	12.52	0.48	0.04
# [Student SHAL - Tutor RS]	15.05	5.38	0.44	0.06
# [Student CA - Tutor POS]	0.15	0.37	−0.51	0.03
# [Tutor DAQ - Student DEEP]	1.70	2.03	0.43	0.07
# [Tutor LAQ - Student DEEP]	2.50	1.79	0.40	0.09
# [Tutor RC - Student DEEP]	2.90	1.65	0.66	0.00
# [Tutor POS - Student SHAL]	43.25	10.78	0.39	0.10
# [Tutor RS - Student SHAL]	13.45	5.36	0.48	0.04

that the best way to use such organizational dialogue acts, which mainly serve to clarify and summarize, is not yet fully understood in our computer tutor.

Table 5 presents our best results on correlations between Bigram Dialogue Act Measures and learning in our human-computer corpus. The grouping of rows now represent the **Student Act**$_n$ **- Tutor Act**$_{n+1}$ results, and the **Tutor Act**$_n$ **- Student Act**$_{n+1}$ results, respectively. Many of these bigrams incorporate, as a first or second element, a unigram from Table 4: student Deep Answers, and tutor Deep Answer Questions and Positive Feedback all positively correlate with learning in isolation, and continue to positively correlate with learning (significantly or as a trend) as part of most larger dialogue act patterns. Moreover, in some cases the positive correlation becomes stronger: e.g., student Deep Answers are more strongly correlated with learning preceding a tutor Short Answer Question or following a tutor Recap than they are in isolation. Interestingly, neither of these tutor Dialogue Acts correlates with learning in isolation. In fact, the ratio of tutor Short Answer Questions is negatively correlated with learning in isolation, suggesting that a

Table 6. *Unigram-learning correlations: human-human corpus (14 students)*

Unigram Dialogue Act Measure	Mean	Std.Dev.	R	p
# Student NOVS	19.29	7.95	0.49	0.09
# Student DEEP	68.50	27.99	−0.49	0.09
Rat: Student NOVS to Answers	0.14	.05	0.47	0.10
Rat: Student SAQ to Question Acts	0.91	0.12	0.56	0.05
Rat: Student LAQ to Question Acts	0.03	.08	−0.57	0.04
# Tutor RD	19.86	10.58	−0.71	0.01
% Tutor RD	5.65	1.83	−0.61	0.03
# Tutor RS	79.14	26.83	−0.56	0.05
# Tutor NEG	14.50	7.60	−0.60	0.03

very specific use of these questions to respond to student Deep Answers may be a better system strategy as compared to their wide-spread use throughout the dialogue.

Table 5 shows a number of Dialogue Acts that *only* correlate with learning in our human-computer corpus when they occur as part of a larger dialogue pattern. For example, tutor Bottom Outs and Expansions are associated with increased student learning only after student Deep Answers, not in isolation, and Tutor Long Answer Questions only correlate with learning after student Shallow Answers or before student Deep Answers. In addition, neither student Shallow Answers nor tutor Restatements correlate in isolation with learning, but there is a positive correlation with learning when the tutor responds to or precedes a student Shallow Answer with a Restatement. Similarly, there is a positive correlation with learning when the tutor responds to or precedes a student Shallow Answer with Positive Feedback, and the strength of the correlation between the [Student SHAL - Tutor POS] bigram is stronger than that of Positive Feedback in isolation. Interestingly, Positive Feedback is *negatively* correlated with learning after student Can't Answers (e.g. "I don't know"). Taken together, these bigram results suggest that generating Positive Feedback in particular dialogue act contexts might be a more effective system strategy than one that doesn't consider the dialogue act context.

5 Human-human results

Table 6 presents our most significant results on correlations between Unigram Dialogue Act Measures and learning in our human-human corpus, using the same format as Table 4. 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 7. *Bigram-learning correlations: human-human corpus (14 students)*

Bigram Dialogue Act Measure	Mean	Std.Dev.	R	p
# [Student NOVS - Tutor BO]	1.50	1.22	0.63	0.02
# [Student DEEP - Tutor DAQ]	8.57	5.61	−0.65	0.02
# [Student DEEP - Tutor LAQ]	2.57	2.38	−0.62	0.02
# [Student DEEP - Tutor NEG]	7.36	5.15	−0.65	0.02
# [Student DEEP - Tutor RS]	17.07	7.81	−0.47	0.10
# [Student DEEP - Tutor RD]	4.21	4.51	−0.63	0.02
# [Student DEEP - Tutor HN]	15.43	10.10	−0.61	0.03
# [Student SHAL - Tutor RS]	18.50	7.59	−0.60	0.03
# [Student SAQ - Tutor POS]	2.07	2.53	-0.52	0.07
# [Student SAQ - Tutor RC]	0.43	0.85	−0.48	0.10
# [Student SAQ - Tutor HN]	1.79	0.98	−0.53	0.07
# [Student LAQ - Tutor DAQ]	0.21	0.43	−0.78	0.00
# [Student LAQ - Tutor POS]	0.14	0.36	−0.71	0.01
# [Student DAQ - Tutor EXP]	0.29	0.61	−0.52	0.07
# [Tutor SAQ - Student DEEP]	13.36	8.43	−0.54	0.06
# [Tutor DAQ - Student SAQ]	0.57	0.85	−0.58	0.04
# [Tutor DAQ - Student LAQ]	0.21	0.43	−0.78	0.00
# [Tutor LAQ - Student SAQ]	0.29	0.61	−0.69	0.01
# [Tutor POS - Student DEEP]	8.57	5.60	−0.64	0.02
# [Tutor POS - Student LAQ]	0.07	0.27	−0.65	0.02
# [Tutor HN - Student SAQ]	1.00	1.18	−0.65	0.02
# [Tutor EX - Student LAQ]	0.14	.53	−0.71	0.01
# [Tutor EX - Student DAQ]	0.07	.27	−0.65	0.02
# [Tutor RS - Student DEEP]	11.36	9.26	−0.64	0.02
# [Tutor RS - Student SHAL]	11.64	6.55	−0.50	0.08
# [Tutor RS - Student SAQ]	1.14	1.41	−0.71	0.01
# [Tutor RD - Student DEEP]	2.07	3.67	−0.68	0.01
# [Tutor RD - Student NOVS]	0.43	0.65	0.57	0.04
# [Tutor RD - Student SHAL]	1.43	1.83	−0.49	0.09

Table 6 also shows that the *type of dialogue acts used by the tutor* relates to how much students learn. 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.

Table 7 presents our best results on correlations between Bigram Dialogue Act Measures and learning in our human-human corpus. Many of these bigrams incorporate, as a first or second element, a unigram from Table 6. In particular, student Novel/Single Answers show a trend to positively correlate with learning in isolation, and continue to positively correlate with learning when followed by a tutor Bottom Out, with the strength of the correlation now reaching significance. Student Deep Answers, and tutor Restatements, Request/Directives and Negative Feedback all negatively correlate in isolation with learning, and also continue to negatively correlate with learning as part of larger dialogue patterns.

Table 7 also shows a number of Dialogue Acts that *only* correlate with learning as part of a larger dialogue pattern. With respect to student acts, although Shallow Answers do not correlate with learning in isolation, they correlate negatively with

learning when they follow or precede a tutor Restatement, or follow a tutor Request/Directive. Moreover, student Questions (Short, Long and Deep) correlate negatively with learning in the context of many tutor acts, even though an increased ratio of student Short Answer Questions positively correlates with learning in isolation. With respect to tutor acts, Deep Answer Questions and Hints relate to decreased student learning as tutor responses to student Deep Answers, and both before and after certain types of student questions, but do not correlate with learning in isolation. More generally, all tutor Questions Act types, which did not correlate with learning in isolation, correlate negatively with student learning when followed by student Deep Answers or Short or Long Questions. Similarly, tutor Positive Feedback, Recaps, and Expansions, which did not correlate in isolation, correlate negatively with learning when preceded or followed by student Deep Answers and/or various student Question Acts.

6 Deep reasoning and correctness

As shown in Section 4, in our computer tutoring corpus, student learning positively correlates both with student utterances displaying reasoning and with tutor questions requiring reasoning, in isolation and as part of larger dialogue patterns. These results are similar to previous human-tutoring findings, where learning correlated both with students' construction of knowledge, and with tutors prompting students to construct knowledge (Chi *et al.* 2001). We hypothesize that because Deep Answers involve more reasoning, they involve more knowledge construction. Note that since we previously found no significant correlation between average turn length (# words/turn) or dialogue length (total words) and learning in our computer- or human-tutoring corpora (Litman *et al.* 2004), this suggests that it is not the quantity but the quality of the students' responses that correlate with learning.

However, in our human tutoring corpus, the relationship between the depth of reasoning displayed in student answers and learning is not so straightforward: while student Novel/Single insights positively correlate with learning, student Deep Answers negatively correlate with learning, alone and as part of larger dialogue patterns. While this negative correlation is surprising, we hypothesize that the *correctness* of student Deep Answers can provide further insight, especially because in the human-human corpus, students speak longer and more freely than in the human-computer corpus, and thus their reasoning may be more error-prone.

We performed a pilot study to investigate this hypothesis. In our human-human corpus, one annotator labeled all the student turns for "Correctness", based on the human tutor's (usually explicit) response to the student answer. Two labels were studied: "Correct", where the tutor considered the answer to be wholly correct, and "Incorrect", where the tutor considered the answer to be wholly or partly incorrect. A second annotator separately labeled a subset of the corpus (507 student turns); inter-annotator agreement was quite high (Kappa = 0.85, "Almost Perfect"). In our human-computer corpus, these same "Correctness" labels are automatically available from the semantic understanding component of ITSPOKE's backend system (Why2-Atlas). For example, in Figure 1, **STUDENT**$_7$ was labeled *Correct* and **STUDENT**$_9$

Table 8. *Deep and correctness – learning correlations: human-human corpus*

Dialogue Act Measure	Mean	Std.Dev.	R	p
% Student Incorrect DEEP	6.85	2.43	−0.52	0.07
# [Student Correct DEEP - Tutor DAQ]	4.71	3.00	−0.70	0.01
# [Student Incorrect DEEP - Tutor LAQ]	1.43	2.03	−0.58	0.04
# [Student Incorrect DEEP - Tutor NEG]	6.50	4.85	−0.58	0.04
# [Student Incorrect DEEP - Tutor RS]	4.29	3.75	−0.52	0.07
# [Student Correct DEEP - Tutor RD]	1.93	2.40	−0.47	0.10
# [Student Incorrect DEEP - Tutor RD]	1.86	2.63	−0.65	0.02
# [Student Incorrect DEEP - Tutor HN]	10.14	8.21	−0.64	0.02
# [Tutor SAQ - Student Incorrect DEEP]	5.64	4.25	−0.61	0.03
# [Tutor POS - Student Correct DEEP]	4.57	2.06	−0.70	0.01
# [Tutor POS - Student Incorrect DEEP]	3.50	3.98	−0.56	0.05
# [Tutor RS - Student Correct DEEP]	6.00	4.30	−0.60	0.03
# [Tutor RS - Student Incorrect DEEP]	4.64	4.14	−0.61	0.03
# [Tutor RD - Student Correct DEEP]	1.21	2.15	−0.61	0.03
# [Tutor RD - Student Incorrect DEEP]	0.79	1.63	−0.67	0.01

was labeled *Incorrect*. In Figure 2, $\mathbf{STUDENT}_{102}$ and $\mathbf{STUDENT}_{106}$ were labeled *Incorrect*, and $\mathbf{STUDENT}_{104}$ was labeled *Correct*.

For each student in our two corpora, we then computed a total for student Correct Deep Answers and student Incorrect Deep Answers, and also a percentage for each, by dividing each total by the total number of student turns. In addition, we recomputed all of the bigrams that contained a student Deep Answer in Tables 5 and 7, above, after incorporating the correctness of the Deep Answer. Effectively, we split each of these original bigrams into two types: one containing student Correct Deep Answers, and one containing student Incorrect Deep Answers. Finally, we re-ran correlations between each of these measures and student learning.

Table 8 presents our significant results and trends on correlations between these measures and learning in our human-human corpus. As shown first, student Incorrect Deep Answers account for the overall negative correlation between student Deep Answers and learning in our human-human corpus; Correct Deep Answers do not correlate (significantly or as a trend) with learning. The second section of the table shows that most of the negatively correlated Student Deep Answer-Tutor Act bigrams in Table 7 are also accounted for by Incorrect Deep Answers. However, the third section of the table shows that most of the negatively correlated Tutor Act-Student Deep Answer bigrams in Table 7 relate to Student Deeps overall, since both Correct and Incorrect Deep Answers continue to negatively correlate.

For comparison, Table 9 presents our significant results and trends on correlations between the same measures and learning in our human-computer corpus. As shown first, student Correct and Incorrect Deep Answers both positively correlate with learning. The second and third sections of the table show that most of our positively correlated bigrams from Table 7 remain positively correlated, either with student Correct Deep Answers or Incorrect Deep Answers, although not both.

Table 9. *Deep and correctness - learning correlations: human-computer corpus*

Dialogue Act Measure	Mean	Std.Dev.	R	p
# Student Correct DEEP	4.05	2.54	0.39	0.10
# Student Incorrect DEEP	7.85	4.02	0.43	0.07
# [Student Correct DEEP - Tutor SAQ]	2.92	2.11	0.41	0.08
# [Student Incorrect DEEP - Tutor DAQ]	1.15	1.14	0.54	0.02
# [Student Incorrect DEEP - Tutor BO]	3.15	2.70	0.46	0.05
# [Student Incorrect DEEP - Tutor EX]	2.80	1.91	0.46	0.05
# [Tutor DAQ - Student Correct DEEP]	0.75	0.91	0.53	0.02
# [Tutor LAQ - Student Incorrect DEEP]	1.40	1.23	0.53	0.02
# [Tutor RC - Student Incorrect DEEP]	2.15	1.27	0.68	0.00

7 Discussion

Overall, our results suggest that student dialogue behaviors, tutor dialogue behaviors, and interacting student and tutor dialogue behaviors all contribute to learning correlations; however, we see little cross-corpora overlap in terms of the specific results. The human-computer correlations were mostly positive, while the human-human correlations were mostly negative, and the specific unigrams and bigrams that yielded correlations were very different across the two corpora. This suggests the importance of designing computational systems using appropriate training data.

Our human-computer results showed that student learning positively correlated with both student Correct and Incorrect Deep Answers, and with tutor Deep Answer Questions, both in isolation and as part of larger dialogue patterns. The bigram analyses showed more generally that student-tutor interaction is an important consideration for learning, since local patterns correlated with learning even when neither bigram element correlated with learning in isolation. In particular, student Shallow Answers correlate with increased learning in the (preceding or following) context of tutor positive feedback or restatements, or when followed by a tutor Long Answer Question that asks for further interpretation or definition. Similarly, tutor Recaps of earlier-established points and tutor requests for definitions or interpretations both correlate with increased learning when followed by student Deep Answers, while tutor Bottom Outs and Expansions are associated with increased learning after student Deep Answers. These results suggest adaptive dialogue strategies that can be implemented in future versions of ITSPOKE.

Our human-human results were more complex. First, although learning positively correlates with student Novel/Single insights both in isolation and even more strongly when followed by tutor Bottom Outs, student Deep Answers negatively correlate with learning both in isolation and as part of larger dialogue patterns. Further investigation showed that the Incorrect Deep Answers largely accounted for these negative correlations; Correct Deep Answers did not correlate with learning significantly or as a trend either in isolation or as part of most Student Deep Answer-Tutor Act bigrams. However, the negative correlations in the Tutor

Act-Student Deep Answer bigrams relate to Student Deeps overall, since both Correct and Incorrect Deep Answers continued to negatively correlate with learning.

Second, while student question-asking is often considered a constructive activity (Chi *et al.* 2001), we did not see a straightforward relation between question-asking and learning: student Short Answer Questions positively correlate with learning, but student Long Answer Questions negatively correlate with learning, in isolation. Moreover, the bigram results show that all student question types negatively correlate with learning in the context of specific tutor dialogue acts. However, there were only 12 Long Answer Questions in our human-human data, and all displayed clear evidence of student misunderstanding (e.g., via phrases such as "what do you mean?"). Finally, tutor State Acts, which attempt to direct the dialogue, are negatively correlated with learning both in isolation and in the context of many student question and answer types. Given the fact that the students in our human-human *are learning*, our results suggest that what is effective in this human tutoring condition is too complicated to be captured in either a unigram or a bigram analysis. In contrast, the human-computer dialogues are much more simple, thus these n-gram techniques are able to capture patterns reflective of increased learning.

8 Conclusions and current directions

This paper presented analyses of student and tutor behavior at the dialogue act level, using unigrams and bigrams of student and tutor dialogue acts to find correlations with student learning in human-human and human-computer spoken tutoring dialogue corpora. The results of our analyses show the importance of analyzing dialogue-learning correlations both from multiple perspectives (student-centered, tutor-centered and interaction-centered), and across different types of tutoring situations (human versus computer tutoring). For example, although we found significant correlations within all of our unigram and bigram measures, the results for specific dialogue acts differed across our corpora, suggesting the importance of training systems from appropriate data. In particular, our human-computer results show that student displays of reasoning, and tutor questions that ask for student reasoning, both positively correlate with learning, alone and within larger dialogue patterns. Our human-human results mirror the greater complexity of human-human interaction: student novel insights positively correlate with learning, but student deeper reasoning (particularly when incorrect) and questioning, as well as tutor attempts to direct the dialogue, all negatively correlate with learning, alone and within larger dialogue patterns. Both corpora also showed that some dialogue acts only correlate with learning as part of larger dialogue patterns.

Our bigrams provide a new empirically-motivated approach for studying the impact of local dialogue interactions on student learning. We plan to gain further insight into our results by investigating correlations between learning and larger n-grams of tutor and student dialogue acts, and also more sophisticated hierarchical structures of dialogue acts (e.g., (Pilkington 1999)). Our future work will also use multiple human tutors to see how our findings generalize; since this is costly, human tutoring studies commonly use one tutor over multiple students (e.g. (Core *et al.*

2003; Rosé *et al.* 2003)). We plan to use the results of our analyses in our system by monitoring and adapting to dialogue act patterns that correlate with learning.

Acknowledgments

This research is supported by ONR (N00014-04-1-0108) and NSF (0325054).

References

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

Core, M. G. and Allen, J. F. (1997) Coding dialogues with the DAMSL annotation scheme. *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pp. 28–35. Menlo Park, CA.

Core, M. G., Moore, J. D. and Zinn, C. (2003) The role of initiative in tutorial dialogue. *Proc. European Chap. Assoc. Computational Linguistics*, Budapest, Hungary.

Forbes-Riley, K., Litman, D., Huettner, A. and Ward, A. (2005) Dialogue-learning correlations in spoken dialogue tutoring. *Proceedings of the International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands.

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

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

Jackson, G., Person, N. and Graesser, A. (2004) Adaptive tutorial dialogue in AutoTutor. *Proc. Workshop on Dialog-based Intelligent Tutoring Systems at ITS'04.*

Katz, S., Allbritton, D. and Connelly, J. (2003) Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence and Education* **13**.

Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–174.

Litman, D. and Silliman, S. (2004) ITSPOKE: An intelligent tutoring spoken dialogue system. *Companion Proc. Human Language Technology: North American Chapter of the Association for Computational Linguistics.*

Litman, D. J., Rose, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D. and Silliman, S. (2004) Spoken versus typed human and computer dialogue tutoring. *Proc. Intelligent Tutoring Systems.*

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

Rickel, J., Lesh, N. B., Rich, C., Sidner, C. L. and Gertner, A. (2001) Building a bridge between intelligent tutoring and collaborative dialogue systems. *Proc. Artificial Intelligence in Education*, pp. 592–594.

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

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

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