Learner characteristics and dialogue: recognising effective and student-adaptive tutorial strategies

Christopher M. Mitchell*, Eun Young Ha, Kristy Elizabeth Boyer and James C. Lester

Department of Computer Science,
North Carolina State University,
890 Oval Dr., Campus Box 8206,
Raleigh, NC 27695, USA
E-mail: cmmitch2@ncsu.edu
E-mail: eha@ncsu.edu
E-mail: keboyer@ncsu.edu
E-mail: lester@ncsu.edu
*Corresponding author

Abstract: In recent years, there have been significant advances in tutoring systems that engage students in rich natural language dialogue. With the goal of further understanding what makes tutorial dialogue successful, this article presents a corpus-based approach to modelling the differential effectiveness of tutorial dialogue strategies with respect to learning. We present results of a study in which task-oriented, textual tutorial dialogue was collected from remote one-on-one human tutoring sessions. This article extends a previous study which found that certain dialogue acts were correlated with learning and student characteristics in the corpus. The predictive models presented here demonstrate important differences between the dialogue sequences that were correlated with learning for different groups of students. The models demonstrate that tutor directives, a type of bottom-out hint, were negatively associated with learning for students with low incoming knowledge or low self-efficacy. The findings signal the importance of tutorial dialogue that adapts to learner characteristics.

Keywords: intelligent tutoring systems; tutorial dialogue; dialogue acts; task-oriented dialogue; learner characteristics; adaptation; learning technology.

Reference to this paper should be made as follows: Mitchell, C.M., Ha, E.Y., Boyer, K.E. and Lester, J.C. (xxxx) 'Learner characteristics and dialogue: recognising effective and student-adaptive tutorial strategies', *Int. J. Learning Technology*, Vol. X, No. Y, pp.000–000.

Biographical notes: Christopher M. Mitchell is a PhD student at North Carolina State University, where he is advised by Kristy Elizabeth Boyer and James C. Lester. He received his MS and BS in Computer Science from North Carolina State University. His research interests lie in computational linguistics, dialogue systems, and intelligent tutoring systems.

Eun Young Ha is a postdoctoral research scholar in the Department of Computer Science at North Carolina State University. Her research interests lie in bringing robust solutions to complex problems at the interface of humans and computers by leveraging and advancing AI to create intelligent interactive

systems. Her research in natural language has focused on discourse processing and natural language dialogue systems, and her research in user modelling has focused on goal recognition.

Kristy Elizabeth Boyer is an Assistant Professor of Computer Science at North Carolina State University. Her research focuses on natural language dialogue, with a particular emphasis on modelling how humans learn through one-on-one interactions. She received her PhD in Computer Science from North Carolina State University, MS in Applied Statistics from the Georgia Institute of Technology, and BS in Mathematics and Computer Science from Valdosta State University. She is the Programme Co-chair for the 2014 International Conference on Intelligent Tutoring Systems.

James C. Lester is Distinguished Professor of Computer Science at North Carolina State University, where he is the Director of the Center for Educational Informatics. His research focuses on transforming education with technology-rich learning environments utilising artificial intelligence, game technologies, and computational linguistics. He has served as the Programme Chair for the International Conference on Intelligent Tutoring Systems, the International Conference on Intelligent User Interfaces, and the International Conference on Foundations of Digital Games, and as an Editor-in-Chief of the *International Journal of Artificial Intelligence in Education*.

This paper is a revised and expanded version of a paper entitled 'Recognizing effective and student-adaptive tutor moves in task-oriented tutorial dialogue' presented at the Intelligent Tutoring Systems Track of the 25th International Conference of the Florida Artificial Intelligence Research Society, Marco Island, Florida, USA, 24 May 2012.

1 Introduction

One-on-one tutoring has been shown to be highly effective (Bloom, 1984; VanLehn, 2010). While the mechanisms that enable such effectiveness are not fully understood, they are explained in part by the rich interactions between students and tutors (Chi et al., 2001), the adaptive presentation of instructional material (D'Mello et al., 2010), motivational strategies (Lepper et al., 1993), and the exchange of rich natural language dialogue (Graesser et al., 1995; Litman et al., 2009). These and other characteristics of natural language tutorial dialogue have been studied extensively in an effort to develop tutorial dialogue systems that are highly effective. Significant progress has been made toward that goal, as evidenced by existing tutorial dialogue systems in diverse domains such as physics (Chi et al., 2010; D'Mello et al., 2011; Forbes-Riley and Litman, 2009), mechanical engineering (Kumar et al., 2010), computer literacy and critical thinking (D'Mello et al., 2011), electricity and electronics (Dzikovska et al., 2010), and computer science (Chen et al., 2011).

Until recently, most prior tutorial dialogue work has proceeded by either implicitly or explicitly assuming that the actions taken most *frequently* by human tutors are the *most effective*. Studying human tutorial dialogue in this way can yield useful insights into the collaborative patterns involved in tutorial dialogue and into the approaches of both expert and non-expert tutors (Boyer et al., 2007; D'Mello et al., 2010). However, it may ultimately be the case that modelling *differential effectiveness*, or which human

tutoring strategies are more effective than others, is the key to building models of tutorial dialogue that approach optimal strategies. For example, one aspect of effective adaptation within tutorial dialogue is to consider learner characteristics such as self-efficacy and incoming knowledge level. It is known that such individual differences influence the structure of tutorial dialogue (D'Mello et al., 2009), and these differences can be used to model adaptations that tutorial dialogue management systems should undertake.

While today's tutorial dialogue systems engage in rich dialogue and effectively support learning, they do not consistently match the effectiveness of expert human tutors for facilitating student learning (VanLehn, 2010). In addition to adapting to learner characteristics as mentioned above, a promising approach for improving the effectiveness of tutorial dialogue systems is to model the association between tutoring strategies and desired outcomes such as learning gains (Chen et al., 2011; Chi et al., 2010; Forbes-Riley and Litman, 2009; Ohlsson et al., 2007). In contrast to earlier work, such an approach does not simply assume that the tutoring strategies used most frequently by humans are the most effective, but rather identifies the most effective strategies by building predictive models of target outcomes based on tutoring strategies.

This article reports findings from predictive modelling in which dialogue profiles of both student and tutor are used within regression analysis to predict learning. Significantly different results emerge based on learner characteristics, indicating the importance of adapting tutorial strategies based on these characteristics. This analysis extends prior work that examined the correlations between dialogue acts and learning effectiveness within the same corpus (Mitchell et al., 2012). This analysis, conducted on a corpus of human-human textual tutorial dialogue for introductory computer science, is part of the larger JavaTutor project to build a tutorial dialogue system that learns its behaviour from corpora with experienced human tutors. The findings reveal correlations between frequencies of dialogue acts and learning outcomes. Additionally, the results reveal ways in which learner characteristics such as self-efficacy and incoming knowledge level are associated with dialogue structure. Finally, we analyse the interaction of learner characteristics with dialogue structure by modelling the differential effectiveness of dialogue acts for different groups of students. The results add to the body of knowledge about ways in which tutorial dialogue is adapted to learner characteristics, and build on prior work that has suggested associations between particular tutoring strategies and learning. These insights will help inform the construction of tutorial dialogue systems that effectively adapt to learner characteristics.

2 Related work

From the early days of tutorial dialogue research, it has been recognised that studying human tutoring is a promising approach to discovering effective strategies that can be utilised within intelligent tutoring systems (Chen et al., 2011; D'Mello et al., 2010; Fox, 1993; Graesser et al., 1995). Because of the proven effectiveness of human tutoring, some work examined the actions of human tutors and adopted the premise that systems ought to employ the strategies that humans employed most frequently. This has been referred to as the 'code-and-count' approach (Ohlsson et al., 2007). Assuming that human tutors' actions are effective can be a reasonable step, particularly when the tutors being studied are highly experienced and have been proven effective over time (Cade et al.,

2008). For example, studying expert human tutors has recently yielded insights into the potential importance of off-topic conversation during tutoring (Lehman et al., 2010), and has suggested ways in which tutors convey information via 'collaborative lecture' (D'Mello et al., 2010).

However, there is growing recognition that human tutors vary in their effectiveness. For example, there is sometimes not a clean distinction between the effectiveness of expert and non-expert tutors (Chen et al., 2011; Cohen et al., 1982; Evens and Michael, 2005). For this reason, it is important to model the differential effectiveness of tutoring approaches – that is, to identify which tutorial dialogue structures are associated with effective learning. In the present article, we examine a corpus that is roughly twice as large as in prior work, learning models of dialogue move effectiveness for separate groups of students based on clustering over a combination of their characteristics.

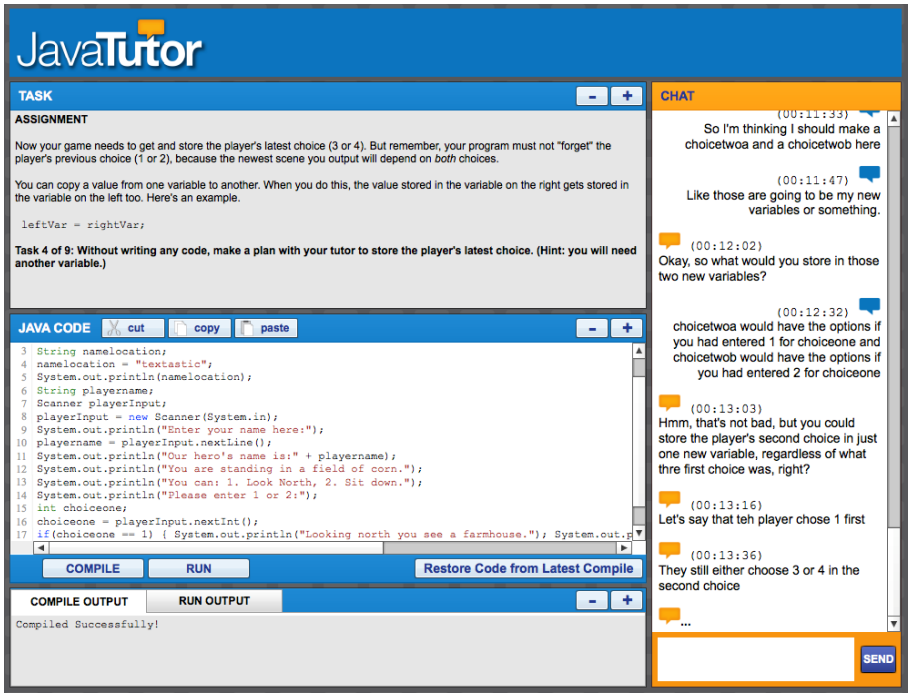
Dialogue has been found to correlate with learning at several levels. These include tutors adapting to uncertainty (Forbes-Riley and Litman, 2009), providing direct procedural instruction (Chen et al., 2011), eliciting information from a student (Chi et al., 2010), and making social dialogue moves when working with a team of tutees (Kumar et al., 2010). Student moves have also been shown to correlate with learning; for example, expressions of disengagement (Forbes-Riley and Litman, 2011) and negative social talk (Dzikovska et al., 2010) may be associated with decreased learning, while student utterances displaying reasoning may be correlated with increased learning (Litman and Forbes-Riley, 2006). This article examines such dialogue phenomena and their relationship with learning in a data-driven fashion, discovering correlations within a multiple regression framework for predicting learning outcomes.

3 Corpus and annotation

The corpus collected for this work consists of human-human tutorial interactions within a web-based remote textual tutoring interface for Java programming. This work is part of the larger JavaTutor project, which aims to create a tutorial dialogue system that learns its behaviour from corpora with experienced human tutors. The present corpus has been the focus of several analyses within the larger project, including automatic discovery of dialogue strategies (Ha et al., 2013; Mitchell et al., 2013) and analysis of student affective states (Grafsgaard et al., 2012, 2013a, 2013b).

The JavaTutor remote tutoring interface (Figure 1) consists of four panes that display the interactive components of the task-oriented tutoring: the current programming task description, the student's Java program, the compilation or execution output associated with the program, and the textual dialogue messages between the student and tutor. The tutor and student interfaces were synchronised in real time. In addition to conversing via textual dialogue with the tutor, the student also modified, compiled, and ran a computer program within the interface. The tutors' actions were constrained to conversing with the student and advancing to the next task, but tutors could see all student actions within the interface in real time. The study reported in this article was the first use of this software, for which it was purpose-built. This software will be made publicly available at the conclusion of the project.

Figure 1 The JavaTutor remote tutoring interface (see online version for colours)



3.1 Study design

The tutoring study paired each student with a tutor for six lessons on introductory Java programming. These sessions were conducted over a period of four weeks, and each session took one hour total, of which forty minutes was allocated to the tutorial dialogue and the remaining time was allocated to completing pre- and post-instruments. The students received full credit for one-half of their semester project in the engineering course in return for their participation. The four tutors were graduate students with prior tutoring experience in Java programming and they were paid for their participation in the study.

The students were selected from a first-year engineering course and were pre-screened to eliminate those with significant self-reported prior programming experience, such as in a formal course on computing. This selection process was used because the JavaTutor project aims to develop a tutoring system for students with no substantial computer science experience. Thus, including students who had completed formal courses in computing would not have been consistent with the goal of observing tutorial dialogue with novices. Among the students who did not report substantial prior computer programming experience and who were therefore included in the study, incoming knowledge was measured with a pre-test as described below.

The design of this observational tutoring study was intended to measure the differential effectiveness of certain dialogue sequences within tutoring; that is, the extent to which tutoring activities were associated with higher or lower learning gains, and not

the differential effectiveness of tutoring versus non-tutoring. For that reason, there was no control condition in which students were only given the tests.

3.2 Data

This article reports on analysis of the first of the six tutoring lessons. We hereafter refer to the corpus of 36 tutor/student dialogues as the *JavaTutor Lesson 1 corpus*. It consists of 4,624 utterances: 3,216 tutor utterances and 1,408 student utterances. An excerpt from the corpus is displayed in Table 1, which includes the dialogue act and task action labels that will be described in detail in the next section. Over the 36 tutoring sessions, the average number of utterances per session was 128.4 (*min* = 74; *max* = 201; *SD* = 32.6). The average number of tutor utterances per session was 89.3 (*min* = 51; *max* = 137; *SD* = 22.8) and the average number of student utterances per session was 39.1 (*min* = 18; *max* = 69; *SD* = 12.5).

Table 1 Dialogue excerpt with dialogue act and task action tags

Tutor: Perfect [POSITIVE FEEDBACK]
Student: [TASK^{0}]
Tutor: OK. Go ahead and test. [DIRECTIVE]
Student: And I don't need anything in the parentheses? [QUESTION]
Tutor: Line 9 is correct. You do NOT need anything inside the parentheses. [ANSWER]
Student: Ok [ACKNOWLEDGEMENT]
Tutor: Good. [POSITIVE FEEDBACK]
Tutor: Moving on. [STATEMENT]
Tutor:[NEXTTASK]
Student:[TASK^{+}]
Student: [TASK^{-}]
Student:[TASK^{+}]
Tutor: Syntactically correct. But there is a logic error [LUKEWARM FEEDBACK]
Tutor: When will the output statement display your request to the player? [QUESTION]
Student: AFTER they put in their name [ANSWER]
Tutor: Exactly [POSITIVE FEEDBACK]

Note: See Tables 3 and 4 for details on tags.

Students completed an identical pretest and posttest for each lesson in order to measure their learning gain. The average pretest score was 52.6% (*min* = 23.5%; *max* = 100%), while the average posttest score was 77.6% (*min* = 41.1%; *max* = 100%). The improvement from pretest to posttest was statistically significant (two-sample paired *t*-test: $p < 10^{-8}$). While the use of identical pretest and posttest likely leads to a practice effect, this practice effect does not inhibit the current work, which is concerned with the *differential* effectiveness of various dialogue sequences. That is, the practice effect would be present across all students, while the differential impact of the individualised tutoring they received will still be captured by the models.

Learning gain was calculated as *posttest-pretest*, and normalised learning gain for each student was calculated as shown in equation (1). This equation includes an

adjustment for non-positive learning gain and for avoiding division by zero due to a perfect pretest score (one occurrence in the current corpus). This formula was derived from the work of Marx and Cummings (2007). The average normalised learning gain for the JavaTutor Lesson 1 corpus was 0.472 (*min* = –0.286; *max* = 1; *SD* = 0.326; *N* = 36).

$$normalisedLearningGain = \begin{cases} \dfrac{posttest - pretest}{1 - pretest}, & posttest > pretest \\ \dfrac{posttest - pretest}{pretest}, & posttest \leq pretest \end{cases} \quad (1)$$

Student characteristics including gender and self-efficacy for computer science were collected via a survey prior to the first tutoring session. Tutors did not have access to any survey or pretest data. Computer science self-efficacy was calculated as the mean of the student's responses to six Likert-scale items (Table 2). These items were adapted from the domain-specific self-efficacy scale (Bandura, 2006). Across the 36 students, the average self-efficacy score was 3.33 out of a possible 5 (*min* = 2.33; *max* = 4.33; *SD* = 0.56).

Table 2 Domain-specific self-efficacy survey questions

Generally I have felt secure about attempting computer programming problems.
I am sure I could do advanced work in computer science.
I am sure that I can learn programming.
I think I could handle more difficult programming problems.
I can get good grades in computer science.
I have a lot of self-confidence when it comes to programming.

3.3 Annotation

A dialogue act annotation protocol was devised and applied to every utterance in the textual dialogue corpus to capture salient events. This annotation scheme was an extension of a prior annotation scheme for task-oriented tutorial dialogue (Boyer et al., 2011). Three human annotators were trained in an iterative process that included collaborative tagging, refinement of the protocol, and independent tagging. A list of the tags in the annotation scheme is shown in Table 3. The textual nature of the dialogue in our corpus likely had an impact on the frequency of certain dialogue acts. Previous work has demonstrated that there are important differences between spoken and textual dialogue in tutoring (Litman et al., 2006). These differences may influence the way dialogue acts are interpreted. For example, in the current study, the number of ACKNOWLEDGMENT dialogue acts may be lower than expected in spoken datasets where 'backchannel' acknowledgments can be used to indicate continued attention. This distinction leads to a slightly more positive interpretation of acknowledgement dialogue moves in the current study than may be the case in spoken dialogue studies.

Dialogue act annotations were independently applied by three trained annotators. During the training process, these annotators achieved pairwise agreement at acceptable levels on independently tagged sessions (Cohen's kappa > 0.8). After training, 5 out of 36 sessions (14% of the corpus) not used during the training process were used to validate the annotation scheme. Each of these sessions was annotated independently by at

least two of the annotators, yielding a Cohen's kappa of 0.79. This agreement level is considered to indicate 'substantial' strength of agreement (Landis and Koch, 1977).

Table 3 The dialogue act annotation scheme

Tag	*Description*	*Frequency*
H (HINT)	Tutor gives advice to help the student proceed with the task	*Tu:* 223 *St:* 0
DIR (DIRECTIVE)	Tutor explicitly tells the student the next step to take; a bottom-out hint	*Tu:* 251 *St*: 0
ACK (ACKNOWLEDGEMENT)	Acknowledgement of a previous utterance; conversational grounding	*Tu:* 67 *St*: 323
RC (REQUEST CONFIRMATION)	Request confirmation or grounding from the other participant	*Tu:* 14 *St*: 1
RF(REQUEST FOR FEEDBACK)	Student requests an assessment of his or her performance or work from the tutor	*Tu:* 0 *St*: 16
PF (POSITIVE FEEDBACK)	Tutor gives a positive assessment of the student's performance	*Tu:* 539 *St*: 0
LF (LUKEWARM FEEDBACK)	Tutor gives an assessment that has both positive and negative elements	*Tu:* 32 *St*: 0
NF (NEGATIVE FEEDBACK)	Tutor gives a negative assessment of the student's performance	*Tu:* 19 *St*: 0
Q (QUESTION)	Related to the task, the session, the educational content, or other non-feedback topics	*Tu:* 632 *St*: 213
A (ANSWER)	Answer to an utterance marked Q	*Tu:* 162 *St*: 549
C (CORRECTION)	Correction/repair of a typo in a previous utterance	*Tu:* 15 *St*: 11
STMT (STATEMENT)	Related to the task, the session, the educational content, or other non-feedback topics	*Tu:* 1256 *St*: 284
O (OTHER)	Other utterances, usually containing only affective content	*Tu:* 6 *St*: 11

Note: *Tu* = tutor, *St* = student.

In addition to the set of dialogue acts that describe student and tutor utterances, we also define a set of task actions that describe the task progress of the student. These actions are shown in Table 4. At the end of each period of student program editing, the edits during that period were compared to the student's final solution using a token-level minimum edit distance algorithm. This algorithm divided the student's program into tokens, as defined by the Java compiler, and measured the minimum number of tokens that needed to be inserted, deleted, or replaced in order to reach the final solution. The

change in this edit distance over a session provided a measure of progress throughout the task. If a period of editing moved the student closer to the final solution, that task action was labelled with the TASK^{+}tag. Conversely, if the edits moved the student farther from the final solution, the TASK^{-} tag was applied. Finally, the TASK^{0} tag indicates that the student's edits to their program did not change the minimum edit distance to the final solution. This 'neutral' task action could occur if a student made an edit that did not change how it executed, such as changing the contents of a string, comment, or variable name. The granularity for applying these tags was determined by an empirical time-based threshold. The threshold of 1.5 seconds of inactivity was arrived at empirically through experimentation with different threshold values ranging from 0.1 seconds to ten seconds. Thresholds shorter than 1.5 seconds resulted in large numbers of task actions in each session, most of which did not represent a true break in task progress. Thresholds longer than 1.5 seconds resulted in orderings of actions that would not be desirable for generalised use in a tutorial dialogue system. For example, a tutor giving positive feedback without any apparent student task action having occurred (in the case of a long threshold) could result in a tutorial dialogue policy that recommends providing positive feedback regardless of the quality of the programming actions the student is engaging in.

Table 4 Task actions

Action	*Description*	*Frequency*
NEXTTASK	The tutor advances the session to the next subtask within a predefined sequence of subtasks	504
TASK^{+}	Positive student task action, a program edit that results in a lower edit distance to the final solution	1,210
TASK^{-}	Negative student task action, a program edit that results in a higher edit distance to the final solution	209
TASK^{0}	Neutral student task action, a program edit that does not change the edit distance to the final solution.	719

4 Dialogue profiles and effective tutor moves

The overall goal of the analysis reported here is to build models of dialogue moves that predict learning and to investigate how the effectiveness of these dialogue moves changes with respect to student characteristics. A model of the differential effectiveness of dialogue acts based on student characteristics is critical when building effective tutorial dialogue systems, and this section presents a stepwise regression approach to extracting such a model. In order to accomplish this, we first used a clustering approach to partition the sessions in the corpus based on similar student characteristics. We then derived a set of features for our regression models based on unigrams and bigrams of dialogue acts and task actions. This section presents the results of the regression models. Detailed discussion of the findings follows in Section 5.

4.1 Clustering by student characteristics

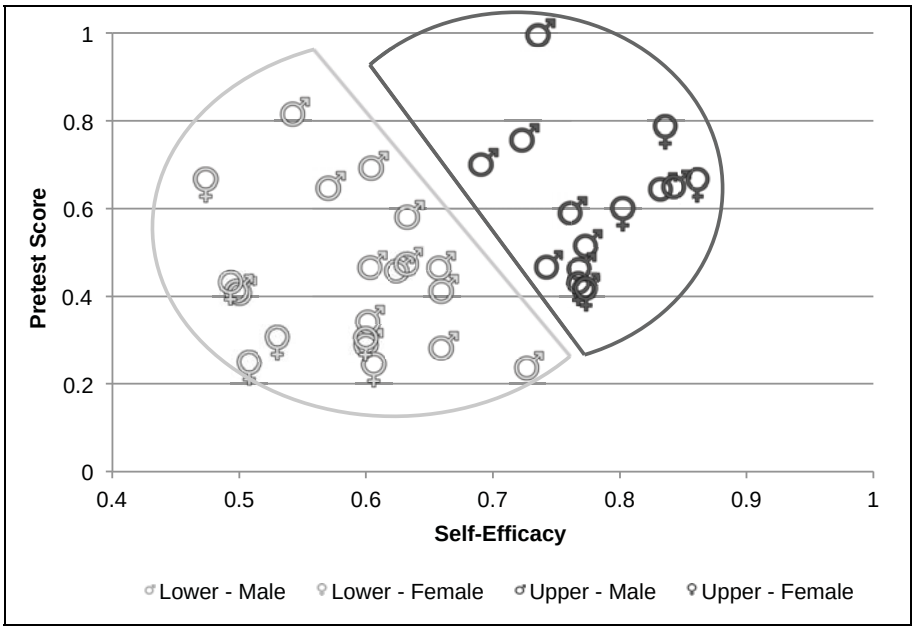
In a previous study with this corpus (Mitchell et al., 2012), it was discovered that there were differences in dialogue profiles between groups of students. For example, those with high incoming knowledge versus low incoming knowledge displayed different relative frequencies of dialogue acts in their tutoring sessions. However, this prior work did not establish differential effectiveness of the strategies observed. In order to model this adaptation formally, the current approach builds separate predictive models of learning for students in different groups in order to model the effectiveness of the adaptive tutorial strategies.

In order to build these separate models by group, one approach could have been to follow prior work and split the data into groups for each student characteristic of interest (e.g., incoming knowledge, self-efficacy) and build a model for each group. However, this approach does not capture dependencies between student characteristics (e.g., in our data most students with high self-efficacy also displayed high incoming knowledge on the pretest). Constructing a model for each combination of each student characteristic would address the limitation regarding the interaction between student characteristics but would result in a number of models equal to the square of the number of student characteristics, which would produce too few data points for training reliable models.

Rather than consider each characteristic individually, we applied a clustering method to identify a meaningful division based on student characteristics. For this analysis, the relevant student characteristics are their incoming domain-specific self-efficacy and their score on the pretest. Clusters were identified using the X-means clustering method, which finds the optimal number of clusters using the Bayesian information criterion[1]. This approach resulted in two clusters, shown in Figure 2. The first cluster, which we will refer to as the *lower* cluster, consists of 21 students with generally lower scores on self-efficacy and pretest. The second cluster, which we will refer to as the *upper* cluster, consists of 15 students with generally higher scores on self-efficacy and pretest. Males and females were approximately evenly split between the two clusters, with approximately 30% of both clusters being female students. The upper cluster had slightly higher normalised learning gains, with an average of 0.52 compared to 0.44 in the lower cluster (this difference was not statistically significant at the $p < 0.05$ level within an independent samples *t*-test). The highest and lowest learning gain sessions were evenly distributed between the two clusters, as seen in Figure 3. So, both clusters contain sessions that were very effective and very ineffective, thus presenting the potential to model variations in learning gains based on dialogue structure.

Figure 4 shows the relative frequency of the most common dialogue acts for each cluster. This graph demonstrates the differences between the clusters in terms of relative frequencies of dialogue acts, showing that tutors and students behaved slightly differently in the two clusters. In particular, Figure 4 shows notable differences in the frequencies of certain dialogue acts between the two clusters. The lower cluster had fewer student ANSWERS and more tutor ANSWERS. The upper cluster had fewer tutor hints. The lower cluster had more student QUESTIONS and fewer tutor QUESTIONS. Finally, the upper cluster had fewer $TASK^0$ events. In the next section we present an analysis of the extent to which these differences between clusters, and moreover, different strategies within clusters, were associated with positive learning outcomes for students.

Figure 2 Pretest and self-efficacy scores for the two clusters used in regression analysis



Note: Points are shifted slightly to reveal overlap.

Figure 3 Histogram of normalised learning gain outcomes for each of the regression clusters

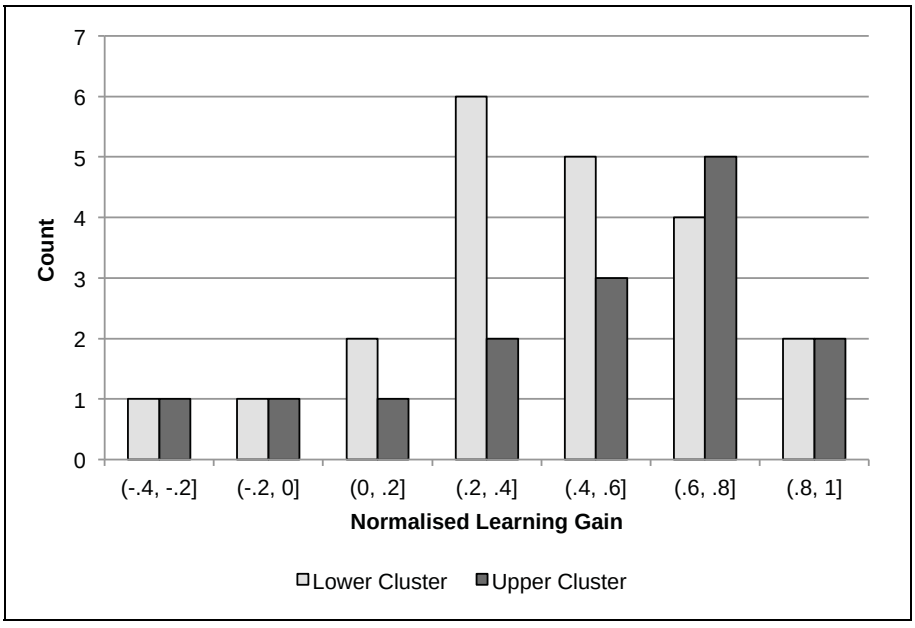
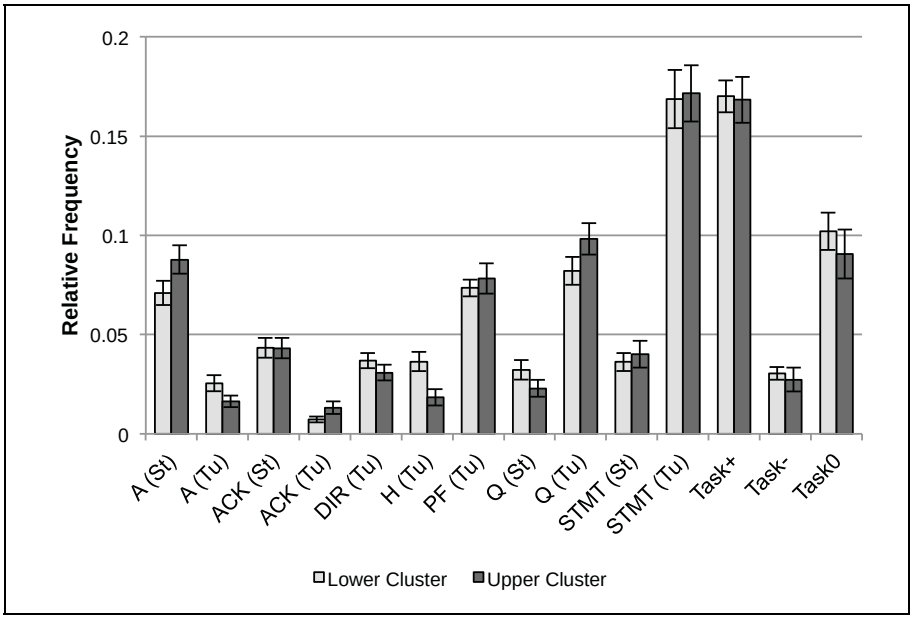


Figure 4 Average relative frequency by cluster for the most common dialogue acts



Notes: Error bars represent standard error from mean. Dialogue act abbreviations are taken from Table 3. St = student, Tu = tutor

4.2 Predicting the effectiveness of tutor adaptations

Our goal was to build a model of the associations between learning outcomes and *moves* in the session, where *moves* include both dialogue acts and task actions. We considered the absolute frequency of single moves, or unigrams, as well as the absolute frequency of pairs of consecutive moves, or bigrams. In order to prevent overfitting in the regression model, we eliminated any unigram or bigram that did not occur in at least two-thirds of the sessions in the subset on which the model was being built, where the three subsets include the following: the overall set that includes all students, the set of students in the lower cluster only, and the set of students in the upper cluster only. This two-thirds threshold was chosen because it partitioned the dialogue acts into those that occurred frequently (in more than 75% of sessions) and those that occurred less frequently (in fewer than 50% of the sessions). No dialogue act occurred with frequency between 50% and 75% in any of the subsets used for this analysis. The subsets on which we built models were the complete dataset (36 sessions), the lower cluster (21 sessions), and the upper cluster (15 sessions). Using this criterion, the same set of 15 unigrams was available for all three subsets. These unigrams are shown in Table 5.

The set of bigrams that occurred in more than two-thirds of sessions differed for each subset, resulting in 37 bigrams in the overall dataset, 42 bigrams in the lower cluster, and 35 bigrams in the upper cluster. Thirty bigrams were shared between all three subsets. In addition to the absolute frequency of unigrams and bigrams, the session length in terms of the total number of moves (dialogue acts + task actions) was included as a possible predictor in the regression models.

Table 5 Dialogue act and task action unigrams used in regression analysis

Student ANSWER	HINT	Student STATEMENT
Tutor ANSWER	POSITIVE FEEDBACK	Tutor STATEMENT
Student ACKNOWLEDGEMENT	NEXTTASK	TASK^{+}
Tutor ACKNOWLEDGEMENT	Student QUESTION	TASK^{-}
DIRECTIVE	Tutor QUESTION	TASK^{0}

The regression models were built using the stepwise linear regression method, which iteratively chooses variables to include and remove from the model, leaving only the variables that surpass a defined significance threshold.[2] A standard significance threshold of $p = 0.05$ was used. Table 6 shows the model for each of the three datasets. In the overall dataset, three bigrams were significantly predictive of learning: (TASK^{+}, TASK^{+}), (TASK^{+}, TUTOR STATEMENT), and (POSITIVE FEEDBACK, DIRECTIVE). In the upper cluster, there were no significant predictors of normalised learning gain. In the lower cluster, there was one unigram, tutor ANSWER, that was significantly negatively associated with learning, along with four bigrams that were negatively associated with learning: (TASK^{+}, TASK^{-}), (TASK^{+}, TUTOR STATEMENT), (DIRECTIVE, TASK^{+}), (POSITIVE FEEDBACK, DIRECTIVE). Also in the lower cluster, three bigrams were significantly positively associated with learning gains: (Student ACKNOWLEDGEMENT, Tutor STATEMENT), (Tutor STATEMENT, TASK^{0}), and (Tutor STATEMENT, tutor QUESTION).

Table 6 Stepwise regression models of normalised learning gain

	Coefficient	*p*	*n*
Overall set, $R^2 = 0.4044$			
TASK^{+}, TASK^{+}	–0.0234	0.0090	465
TASK^{+}, Tutor STATEMENT	–0.0560	0.0374	79
POSITIVE FEEDBACK, DIRECTIVE	–0.1378	0.0072	40
Lower cluster (low self-efficacy, low pretest), $R^2 = 0.9429$			
Tutor ANSWER	–0.0479	0.0013	115
Student ACKNOWLEDGEMENT, Tutor STATEMENT	0.0399	0.0393	58
TASK^{+}, TASK^{-}	–0.1924	<0.0001	34
TASK^{+}, Tutor STATEMENT	–0.1112	<0.0001	43
DIRECTIVE, TASK^{+}	–0.1311	<0.0001	29
POSITIVE FEEDBACK, DIRECTIVE	–0.0687	0.0290	27
Tutor STATEMENT, TASK^{0}	0.1227	<0.0001	28
Tutor STATEMENT, Tutor QUESTION	0.0287	0.0103	112
Upper cluster (high self-efficacy, high pretest), $R^2 = 0$			
No significant predictors found			

Note: n = total occurrences in dataset.

5 Discussion

The regression models indicate important relationships between dialogue profiles and learning. The output of the stepwise regression algorithm demonstrated that the frequencies of several dialogue act unigrams and bigrams were associated with learning outcomes in this corpus. In addition, the differences between the regressions for the different subsets of the corpus suggest that student characteristics played a role in the effectiveness of certain dialogue acts. In this section, we interpret the results for each subset of the corpus and discuss the limitations of our approach.

5.1 Dialogue structure and learning: all students

In the overall dataset which includes all students, the first of the significantly predictive bigrams was ($TASK^+$, $TASK^+$), representing two positive task actions in a row. Increased frequency of this bigram was negatively associated with learning gain. This finding may initially seem counterintuitive, since a student's consistent progress would normally be considered indicative of her having learned the material well. However, there are two factors that shed light on this finding. The first is the manner in which task actions were segmented. As described in Section 3.3, the empirically defined threshold is to define a task action as ending after 1.5 seconds of inactivity from the student. Thus, two positive task actions in a row could indicate fragmented progress towards the goal. Additionally, a high occurrence of this bigram indicates that a student made progress without any feedback or other interventions from the tutor. This lack of immediate feedback may not be the most effective approach for complex tasks such as programming, as observed in the literature (Shute, 2008). This interpretation is consistent with other recent analysis of this corpus, which found that tutor intervention during problem solving was preferable to non-intervention (Mitchell et al., 2013).

The second bigram significant in the model, ($TASK^+$, Tutor STATEMENT), is also negatively correlated with learning gains. Tutor STATEMENTS in this tagging scheme do not provide feedback or directly address making progress on the task; those goals are addressed with feedback dialogue acts. Therefore, this negative association could indicate that statements, which provide additional information but no direct feedback on the task, may not be optimal immediately following task progress. An example of this bigram is shown in Table 7, Excerpt A, in which the student undertook two task actions that moved him closer to the final solution, but rather than providing direct task feedback, the tutor provided indirect feedback statements that may have been difficult for a novice to link back to the task at hand.

The last bigram, (POSITIVE FEEDBACK, DIRECTIVE), is also negatively correlated with learning. Upon inspecting instances of this bigram in the corpus, we found that these DIRECTIVES were most often a tutor reminding a student to test her program by compiling and running it, as shown in Table 7, Excerpt B. Although there are several possible explanations for the negative association between this bigram and learning, it is important to note that the concept of testing a program was emphasised in the lesson. Therefore, if a student required reminding of this process regularly, this could be symptomatic of the student not internalising the concepts, which could naturally lead to lower learning gain. While the (POSITIVE FEEDBACK, DIRECTIVE) bigram can be viewed in the aforementioned way as a symptom of the causal problem, it is also possible that this tutorial strategy could be causal in itself. Directing the student more frequently

could have undermined the student's sense of control over the problem-solving process, particularly just after positive feedback (Lepper et al., 1993).

Table 7 Excerpts from the corpus including bigrams from Table 6

Excerpt A	Tutor: Okay, feel free to write the variable declaration with such a name [STATEMENT]
(TASK^{+}, Tutor STATEMENT)	Student: [TASK^{+}]
	Tutor: Note that there are two parts to the declaration statement, the data type and the name of the variable [STATEMENT]
	Student: [TASK^{+}]
	Tutor: Here, you've just made a string variable [STATEMENT]
Excerpt B	Tutor: [NextTASK]
(TASK^{+}, TASK^{+})	Student: [TASK^{+}]
(POSITIVE FEEDBACK, DIRECTIVE)	Student: [TASK^{+}]
	Student: [TASK^{+}]
	Student: [TASK^{+}]
	Student: [TASK^{+}]
	Tutor: Looks good [POSITIVE FEEDBACK]
	Tutor: copile and run it. [DIRECTIVE]
Excerpt C	Tutor: For String, I think it would be good if you remembered it as "declaring a String variable, which can hold a sequence of characters" [STATEMENT]
(Student ACKNOWLEDGEMENT, Tutor STATEMENT)	Student: ah okay [ACKNOWLEDGEMENT]
	Tutor: There are a number of different data types you'll find throughout these lessons [STATEMENT]
Excerpt D	Student: [TASK^{+}]
(Tutor STATEMENT, TASK^{0})	Tutor: Java programmers have a convention of starting variable names with lowercase [STATEMENT]
	Tutor: So, we'd usually write myGame [STATEMENT]
	Student: [TASK^{0}]
	Tutor: Exactly [POSITIVE FEEDBACK]
Excerpt E	Tutor: When you put your variable in the output statement, it displayed the value stored inside the variable. [STATEMENT]
(Tutor STATEMENT, Tutor QUESTION)	Tutor: Is that what you expected? [QUESTION]
	Student: kind of [ANSWER]

Note: Typographical errors originated in corpus.

5.2 Dialogue structure and learning: 'lower' cluster

The results from the lower cluster reveal numerous associations between dialogue structure and learning outcomes. Two bigrams from the overall model reappear here: (TASK^{+}, Tutor STATEMENT) and (POSITIVE FEEDBACK, DIRECTIVE), both having

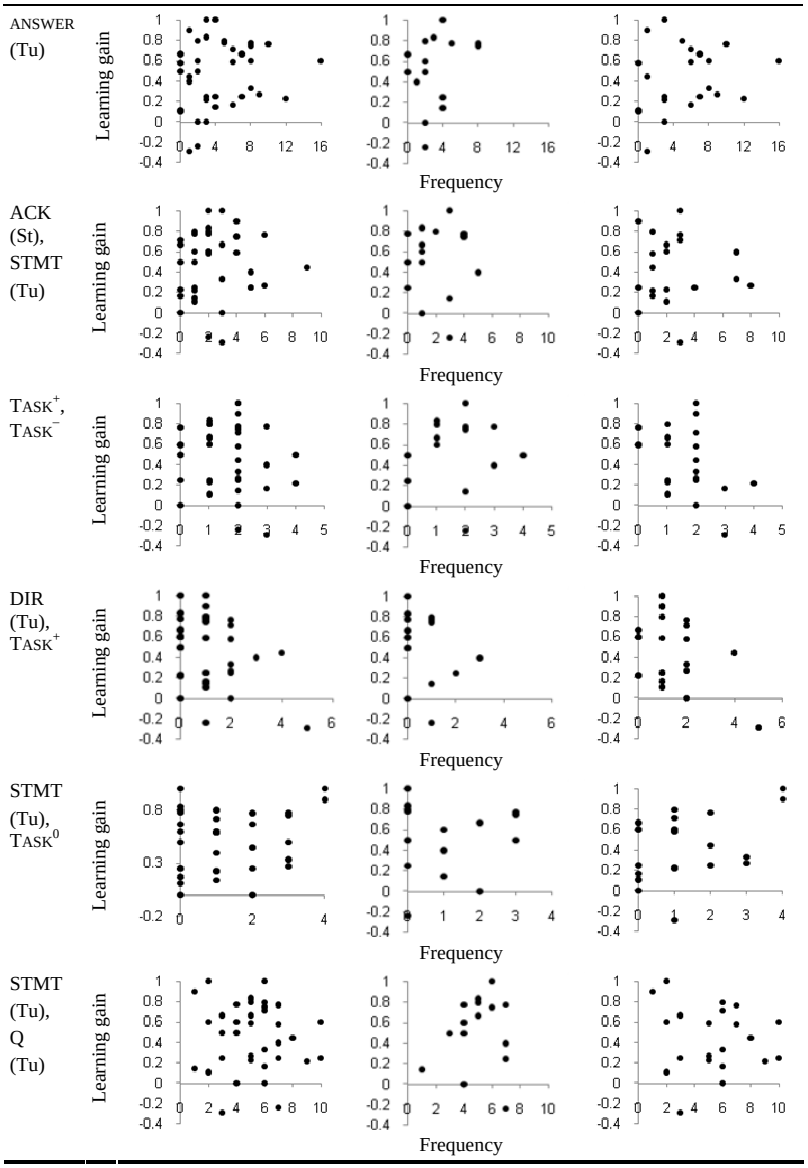
negative correlations with learning gain once again in this model. In addition to these features, the model for the lower cluster identified several other significant predictors. To provide insight into why these predictors were chosen for the model of the lower cluster and not for the models of the upper cluster or the overall dataset, Figure 5 shows scatter plots of the frequency of each significant predictor against the normalised learning gain for each of the three subsets of the corpus used in this analysis.

As shown in the figure, one bigram that was not present in the overall model was the ($TASK^{+}$, $TASK^{-}$) bigram, which had a negative correlation with learning gain in the lower cluster. This may be for the same reason as the ($TASK^{+}$, $TASK^{+}$) bigram in the overall cluster: it represents a fragmented set of task progress events during which there was no tutor intervention. Because the stepwise linear regression method adds the variable with the greatest explanatory power at each step, and because a new variable is only added if it provides significant additional explanatory power, it is possible for different variables representing a similar phenomenon to be chosen in the models for each subset of the corpus. The ($TASK^{+}$, $TASK^{+}$) bigram is shown in Table 7, Excerpt B.

A tutor DIRECTIVE followed by a $TASK^{+}$ action was also negatively correlated with learning gain. By following a tutor's instructions to accomplish the task goal, a student was not required to independently plan the problem-solving steps, and may also have experienced a lower level of control over the problem-solving process. This observation is consistent with our prior finding that DIRECTIVES and bigrams of two directives were significantly negatively correlated with learning gain (Mitchell et al., 2012). It is also possible that directives are symptomatic of students' lack of progress, leading to diminished learning gains.

The final negative correlation with learning gain, the unigram frequency of tutor ANSWERS, is more difficult to explain. We hypothesised that the phenomenon of interest could actually be unanswered student questions, for which the frequency of tutor ANSWERS was serving as a proxy within the models. To investigate further, we computed the difference between the number of student questions and the number of tutor answers in each session within this cluster in order to measure the number of student questions that went unanswered. In just under half of the sessions, the tutor provided at least one answer for every student question. The remaining sessions ranged from a single unanswered student question up to nine unanswered questions. The sessions with unanswered questions had a higher average learning gain, at 0.494, versus the average of 0.366 for the sessions that did not include unanswered questions. Although this difference is not statistically significant, it provides some insight into the correlation in the regression model. It is possible that not directly answering student questions was an effective tutorial strategy within this context, in that it allowed for students to generate their own answers, as in the Socratic method of instruction (Rosé et al., 2001).

Figure 5 Scatter plots of dialogue act unigram and bigram frequencies against normalised learning gain for each subset of the corpus



Notes: Includes significant predictors within regression models for the lower cluster only. Dialogue act abbreviations are taken from Table 3. St = student, Tu = tutor.

Because of the importance of questions and answers within the model, we also investigated whether students who asked more questions tended to learn more. Because the frequency of student questions did not appear in the model, this difference was not expected to be statistically significant. To test the hypothesis, we split students in the cluster based on whether they asked more or fewer questions than the median frequency. We found that students who asked more questions tended to achieve higher learning gain than those who asked fewer questions (0.538 vs. 0.278, $p = 0.053$). This p-value indicates a trend not at the threshold for inclusion within the regression model. However, this positive relationship between learning and student questions, when taken in context with the negative relationship between learning and number of tutor answers, suggests that selectively answering student questions was an effective strategy within the corpus. Investigation into the types of questions that were not answered, or the situations in which they were not answered, is a topic for future work.

There were three bigrams that were positively correlated with learning in the lower cluster. The first of these was a student ACKNOWLEDGEMENT followed by a tutor STATEMENT. These pairs occurred often when a tutor was explaining a concept to a student. The student would acknowledge her understanding of the concept being explained, and the tutor would then expand on the explanation with another statement. For an example of this, see Table 7, Excerpt C. In these situations, the tutor is reinforcing what the student has just learned with additional information about the concept, or by summarising what has just been learned.

The second significant bigram, a tutor STATEMENT followed by a TASK^0 action, often occurred when a tutor gave non-essential advice to a student; for example, on naming conventions for variables or on string formatting, which are important for style but not for fulfilling the task requirements. An example of this is shown in Table 7, Excerpt D, in which the student selected a variable name that was not in keeping with typical conventions of the programming language. The tutor gave a statement that constituted indirect feedback, and the student then made the suggested change. The variable name change was categorised as a TASK^0 action by the edit distance algorithm because it did not have an effect on how the program operated. After this variable name edit, the tutor provided positive feedback. This example illustrates that TASK^0 actions are pedagogically interesting because they indicate that the tutor felt the student was progressing sufficiently well as to offer advice on further improving the computer program beyond what was required.

Finally, a tutor STATEMENT followed up with a QUESTION, as in Excerpt E of Table 7, was positively correlated with learning. In this excerpt the tutor provides a brief explanation for the program behaviour that the student had observed and then asks the student to reflect on this explanation. Another common situation in which this bigram occurred was when a tutor had finished explaining a concept and was prompting the student for the next step to take in the task or testing the student's knowledge of what had just been explained. In this context, it seems that the goal of these tutor questions is to gauge the student's understanding.

5.3 Dialogue structure and learning: 'upper' cluster

The stepwise regression procedure for the upper cluster identified no significant predictors for normalised learning gain. One possible reason for a model to identify no significant predictors would be if the range of the response variable were limited.

However, this is not the case; the range of normalised learning gain (the response variable) for the upper cluster was –0.235 to 1 (*SD* = 0.336), indicating similar spread as the overall cluster, which had normalised learning gains ranging from –0.286 to 1 (*SD* = 0.326). The size of this cluster (N_{upper} = 15 compared with N_{lower} = 21) may play a role in the absence of significant predictors, but it is also possible that the students in the upper cluster, having high self-efficacy and higher pretest scores, were more resilient to suboptimal tutoring strategies than the students in the lower cluster. That is, these students were able to learn more independently and thus were not as reliant on the tutor for help in accomplishing the given tasks. Indeed, the combined number of moves (dialogue acts + task actions) was significantly lower for the upper cluster on average compared to the lower cluster (187 vs. 212, p = 0.048, two-tailed t-test), indicating decreased interaction between students and tutors in the upper cluster.

5.4 Limitations

These results provide insight into the effectiveness of particular tutor moves in a task-oriented dialogue, which holds promise for devising tutorial strategies that adapt based on learner characteristics. However, the current work has several limitations. One of the limitations stems from the annotations themselves: for example, the TASK^0 event is treated by the edit distance algorithm as a neutral edit to the computer program. However, it often preceded tutor POSITIVE FEEDBACK, as seen in Table 7, Excerpt D. Expanding the task annotation to include these pedagogically relevant task events in a separate category can provide the model with additional context for modelling strategies. In addition, a finer-grained breakdown of high frequency dialogue acts, such as POSITIVE FEEDBACK, QUESTIONS, and STATEMENTS would provide more insight into the aspects of a particular tutor utterance that make it effective for learning.

6 Conclusions

Building an automated tutor with the same or greater effectiveness of an expert human tutor is a long-standing goal of intelligent tutoring systems research. A highly promising approach is to model the ways in which human tutorial dialogue moves are correlated with learning, and to subsequently implement the most effective strategies within an intelligent system. This article has described the collection and annotation of the JavaTutor corpus of textual task-oriented tutorial dialogue, and has presented predictive models of learning outcomes based on dialogue structure features. The findings revealed ways in which several tutorial dialogue events differ in their effectiveness for students with different characteristics. These findings pave the way for confirmatory future investigations across tutorial domains, as well as considering larger populations of students from a variety of academic backgrounds.

In addition to future confirmatory studies capable of establishing causality, important directions for future work include investigating the impact of the affective properties of tutor and student dialogue on learning gains. This direction holds particular promise when a tutor and student interact repeatedly, allowing the tutor to form a long-term student model and establishing a robust rapport. Finally, the strategies discovered in this and other analyses of human tutoring must be tested in the context of human-computer tutoring. It is hoped that these lines of investigation will enable the creation of highly

effective tutorial dialogue systems by modelling the differential effectiveness of tutoring strategies, and enabling fine-grained adaptation to learner characteristics.

Acknowledgements

The authors wish to acknowledge the efforts of everyone who has contributed to the JavaTutor project, including Joseph Grafsgaard, Alok Baikadi, Megan Hardy, Bradford Mott, Mary Luong, Miles Smaxwell, Natalie Kerby, Robert Fulton, Caitlin Foster, Denae Ford, Joseph Wiggins, and Robert Hudson. This work is supported in part by the National Science Foundation through Grants DRL-1007962 and CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

Bandura, A. (2006) 'Guide for constructing self-efficacy scales', in F. Pajares and T. Urdan (Eds.): *Self-Efficacy Beliefs of Adolescents*, pp.307–337, Information Age Publishing, Greenwich, Connecticut.

Bloom, B. (1984) 'The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring', *Educational Researcher*, Vol. 13, No. 6, pp.4–16.

Boyer, K.E., Phillips, R., Ingram, A., Ha, E.Y., Wallis, M., Vouk, M. and Lester, J. (2011) 'Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach', *International Journal of Artificial Intelligence in Education*, Vol. 21, No. 1, pp.65–81.

Boyer, K.E., Vouk, M.A. and Lester, J.C. (2007) 'The influence of learner characteristics on task-oriented tutorial dialogue', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.365–372, Marina Del Rey, California.

Cade, W., Copeland, J. and Person, N. (2008) 'Dialogue modes in expert tutoring', in *Proceedings of the International Conference on Intelligent Tutoring Systems*, pp.470–479, Montréal, Canada.

Chen, L., Di Eugenio, B., Fossati, D., Ohlsson, S. and Cosejo, D. (2011) 'Exploring effective dialogue act sequences in one-on-one computer science tutoring dialogues', in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pp.65–75, Portland, Oregon.

Chi, M., VanLehn, K. and Litman, D. (2010) 'Do micro-level tutorial decisions matter: applying reinforcement learning to induce pedagogical tutorial tactics', in *Proceedings of the International Conference on Intelligent Tutoring Systems*, pp.224–234, Pittsburgh, Pennsylvania.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T. and Hausmann, R.G. (2001) 'Learning from human tutoring', *Cognitive Science*, Vol. 25, No. 4, pp.471–533.

Cohen, P.A., Kulik, J.A. and Kulik, C-L.C. (1982) 'Educational outcomes of tutoring: a meta-analysis of findings', *American Educational Research Journal*, Vol. 19, No. 2, pp.237–248.

D'Mello, S.K., Hays, P., Williams, C., Cade, W., Brown, J. and Olney, A. (2010) 'Collaborative lecturing by human and computer tutors', in *Proceedings of the International Conference on Intelligent Tutoring Systems*, pp.178–187, Pittsburgh, Pennsylvania.

D'Mello, S.K., Lehman, B. and Graesser, A.C. (2011) 'A motivationally supportive affect-sensitive AutoTutor', in Calvo, R.A. and D'Mello, S.K. (Eds.): *New Perspectives on Affect and Learning Technologies*, pp.113–126, Springer, New York.

D'Mello, S.K., Olney, A. and Person, N. (2010) 'Mining collaborative patterns in tutorial dialogues', *Journal of Educational Data Mining*, Vol. 2, No. 1, pp.1–37.

D'Mello, S.K., Williams, C., Hays, P. and Olney, A. (2009) 'Individual differences as predictors of learning and engagement', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp.308–313, Austin, Texas.

Dzikovska, M., Steinhauser, N., Moore, J.D., Campbell, G.E., Harrison, K.M. and Taylor, L.S. (2010) 'Content, social, and metacognitive statements: an empirical study comparing human-human and human-computer tutorial dialogue', in *Proceedings of the European Conference on Technology Enhanced Learning*, pp.93–108, Barcelona, Spain.

Evens, M.W. and Michael, J. (2005) *One-on-One Tutoring by Humans and Computers*, Erlbaum, Mahwah, New Jersey.

Forbes-Riley, K. and Litman, D. (2009) 'Adapting to student uncertainty improves tutoring dialogues', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.33–40, Brighton, UK.

Forbes-Riley, K. and Litman, D. (2011) 'When does disengagement correlate with learning in spoken dialog computer tutoring?', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.81–89, Auckland, New Zealand.

Fox, B.A. (1993) *The Human Tutorial Dialogue Project: Issues in the Design of Instructional Systems*, Erlbaum, Hillsdale, New Jersey.

Graesser, A.C., Person, N. and Magliano, J.P. (1995) 'Collaborative dialogue patterns in naturalistic one-to-one tutoring', *Applied Cognitive Psychology*, Vol. 9, No. 6, pp.495–522.

Grafsgaard, J.F., Fulton, R.M., Boyer, K.E., Wiebe, E.N. and Lester, J.C. (2012) 'Multimodal analysis of the implicit affective channel in computer-mediated textual communication', in *Proceedings of the ACM International Conference on Multimodal Interaction*, pp.145–152, Santa Monica, California.

Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J.C. (2013a) 'Automatically recognizing facial indicators of frustration: a learning-centric analysis', in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pp.159–165, Geneva, Switzerland.

Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N. and Lester, J.C. (2013b) 'Embodied affect in tutorial dialogue: student gesture and posture', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.1–10, Memphis, Tennessee.

Ha, E.H., Mitchell, C.M., Boyer, K.E. and Lester, J.C. (2013) 'Learning dialogue management models for task-oriented dialogue with multiple communicative channels', in *Proceedings of the SIGDIAL Meeting on Discourse and Dialogue*, pp.204–213, Metz, France.

Kumar, R., Ai, H., Beuth, J. and Rosé, C.P. (2010) 'Socially capable conversational tutors can be effective in collaborative learning situations', in *Proceedings of the International Conference on Intelligent Tutoring Systems*, pp.156–164, Pittsburgh, Pennsylvania.

Landis, J.R. and Koch, G.G. (1977) 'The measurement of observer agreement for categorical data', *Biometrics*, Vol. 33, No. 1, pp.159–174.

Lehman, B., Cade, W. and Olney, A. (2010) 'Off topic conversation in expert tutoring: waste of time or learning opportunity?', in *Proceedings of the International Conference on Educational Data Mining*, pp.101–110, Pittsburgh, Pennsylvania.

Lepper, M.R., Woolverton, M., Mumme, D.L. and Gurtner, J-L. (1993) 'Motivational techniques of expert human tutors: lessons for the design of computer-based tutors', in Lajoie, S.P. and Derry, S.J. (Eds.): *Computers as Cognitive Tools*, pp.75–105, Erlbaum, Hillsdale, New Jersey.

Litman, D. and Forbes-Riley, K. (2006) 'Correlations between dialogue acts and learning in spoken tutoring dialogues', *Natural Language Engineering*, Vol. 12 No. 2, pp.161–176.

Litman, D., Moore, J.D., Dzikovska, M. and Farrow, E. (2009) 'Using natural language processing to analyze tutorial dialogue corpora across domains and modalities', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.149–156, Brighton, UK.

Litman, D., Rosé, C., Forbes-Riley, K., VanLehn, K., Bhembe, D. and Silliman, S. (2006) 'Spoken versus typed human and computer dialogue', *International Journal of Artificial Intelligence in Education*, Vol. 16, No. 2, pp.145–170.

Marx, J.D. and Cummings, K. (2007) 'Normalized change', *American Journal of Physics*, Vol. 75, No. 1, pp.87–91.

Mitchell, C.M., Boyer, K.E. and Lester, J.C. (2013) 'A Markov decision process model of tutorial intervention in task-oriented dialogue', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.828–831, Memphis, Tennessee.

Mitchell, C.M., Ha, E.H., Boyer, K.E. and Lester, J.C. (2012) 'Recognizing effective and student-adaptive tutor moves in task-oriented tutorial dialogue', *in Proceedings of the Intelligent Tutoring Systems Track of the International Conference of the Florida Artificial Intelligence Research Society*, pp.450–455, Marco Island, Florida.

Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X. and Kershaw, T.C. (2007) 'Beyond the code-and-count analysis of tutoring dialogues', in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp.349–356, Marina Del Rey, California.

Rosé, C.P., Moore, J.D., VanLehn, K. and Allbritton, D. (2001) 'A comparative evaluation of Socratic versus Didactic tutoring', in *Proceedings of the Annual Conference of the Cognitive Science Society*, Edinburgh, Scotland.

Shute, V.J. (2008) 'Focus on formative feedback', *Review of Educational Research*, Vol. 78, No. 1, pp.153–189.

VanLehn, K. (2011) 'The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems', *Educational Psychologist*, Vol. 46, No. 4, pp.197–221.

Notes

1 The Weka implementation was used for this work (http: //www.cs.waikato.ac.nz/ml/weka/).

2 The SAS implementation was used for this work (http: //www.sas.com).