Identifying Effective Moves in Tutoring: On the Refinement of Dialogue Act Annotation Schemes

Alexandria Katarina Vail and Kristy Elizabeth Boyer

Department of Computer Science
North Carolina State University
Raleigh, North Carolina, USA
{akvail,keboyer}@ncsu.edu

Abstract. The rich natural language dialogue that is exchanged between tutors and students has inspired many successful lines of research on tutorial dialogue systems. Yet, today's tutorial dialogue systems do not regularly achieve the same level of student learning gain as has been observed with expert human tutors. Implementing models directly informed by, and even machine-learned from, human-human tutorial dialogue is highly promising. With this goal in mind, this paper makes two contributions to tutorial dialogue systems research. First, it presents a dialogue act annotation scheme that is designed specifically to address a common weakness within dialogue act tag sets, namely, their dominance by a single large majority dialogue act class. Second, using this new fine-grained annotation scheme, the paper describes important correlations uncovered between tutor dialogue acts and student learning gain within a corpus of tutorial dialogue for introductory computer science. These findings can inform the design of future tutorial dialogue systems by suggesting ways in which systems can adapt at a fine-grained level to student actions.

1 Introduction

It has been widely demonstrated that one-on-one tutoring is more effective than many other forms of instruction [1, 2]. This success is thought to be largely a result of the rich natural language interaction between student and tutor [3–5]. Human tutorial dialogue has therefore been studied extensively, and the strategies observed with human tutors have inspired a number of successful tutorial dialogue systems (e.g., [6–10]). However, despite the rapid progress achieved in modern tutorial dialogue systems, systems do not yet match the effectiveness of expert human tutors [1]. A promising direction for further improving tutorial dialogue systems is to identify direct associations between measured student learning gain and tutorial strategies [6, 10–12].

Tutorial strategies are realized at the level of *dialogue acts*, which characterize the intent of dialogue utterances. This paper explores the dialogue acts that human tutors make and identifies relationships between particular dialogue events

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 199–209, 2014.

and student learning gain. The analyses were conducted on a corpus of human-human textual dialogue collected through a tutoring interface for introductory computer science. This study is part of the larger JavaTutor project that is developing an intelligent tutoring system whose behaviors are machine-learned from corpora of human-human tutoring. This paper makes two novel contributions. First, it presents a tutorial dialogue act annotation scheme that addresses an important weakness of prior annotation schemes applied in numerous tutorial dialogue domains: the presence of a large majority class dialogue act that is more vague than other acts and that presents challenges for machine-learning models. Second, this paper utilizes a corpus manually tagged with this refined dialogue act tag set to explore relationships between dialogue acts and student learning gain at the end of the tutoring session. The results suggest important relationships between tutor choices and student learning.

2 Related Work

It has long been recognized that one-on-one tutoring is one of the more effective methods of instruction [13] and that the study of human-human tutorial interactions is crucial to the development of effective intelligent tutorial systems addressing this need [5]. Several dialogue acts have been previously identified as significantly correlated with learning gain [10]; in particular, specific collaborative acts between tutor and student have been studied and established as influential [14]. Historically, it was often assumed that the most frequent human tutorial acts are the effective tutorial acts, since human tutors are considered to be generally effective [11]. This might not be the best approach, as effective tutorial strategies vary from student to student and tutor to tutor [10].

Moving beyond this pure frequency approach, dialogue has been demonstrated to correlate with learning gain in a variety of ways: particular dialogue act sequence occurrences [10], adaptation to dialogue structure correlated with positive learning gain [6], and responsiveness to student uncertainty [12]. However, a frequent limitation in capturing dialogue acts for tutoring and across a variety of dialogue domains lies with crafting the annotation scheme, where it is often discovered after annotation that one dialogue act encompasses a larger portion of the corpus than any other act. For example, the Inform tag comprises 29% of an airline reservation human-human dialogue corpus [15], and the Non-Substantive Act tag, defined to be any act that was not a question, feedback, or answer, comprises 46% of an ITSPOKE physics tutorial dialogue corpus [16].

This paper expands upon prior work by defining a novel annotation scheme derived from a variety of prior schemes. With this refined annotation scheme, the analysis produced new statistical relationships not previously identified between student learning gain and the dialogue events of the tutoring sessions.

3 Tutorial Dialogue Corpus

The corpus examined here consists of computer-mediated textual human-human interactions. The sessions were conducted within an online remote tutoring

interface for Java programming. The interface, displayed in Figure 1, consists of four panes: the task description, the compilation and execution output, the student's Java source code, and the textual dialogue messages between the tutor and the student. The content of the interface was synchronized in real time between the student and the tutor; however, the tutor's interactions with the environment were constrained to the textual dialogue with the student and the ability to progress between tasks.

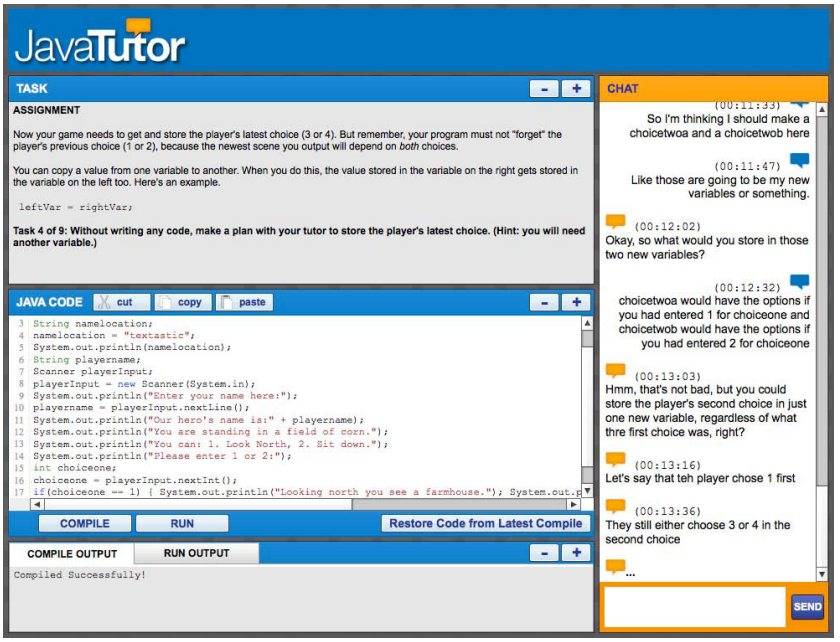

Fig. 1. The tutorial dialogue interface

The full tutorial dialogue corpus under consideration was collected in Fall 2011 and Spring 2012. Due to the time requirements of manual annotation, the current analysis examines a subset of the full corpus, sessions between 30 students paired with one of five tutors for the first of six sequential lessons [17]. This 30-session corpus consists of 4,035 utterances: 2,846 (71%) tutor utterances and 1,189 (29%) student utterances. The average number of utterances per tutorial session is 134.5 (min = 69, max = 213); tutors averaged 94.9 utterances per session (min = 46, max = 159), and students 39.6 utterances per session (min = 21, max = 65). Many utterances contained multiple dialogue acts; to address this concern, the utterances were manually partitioned at sentential and phrasal boundaries by the principal dialogue act annotator. Two sample excerpts from the corpus after annotation are displayed in Figure 2. (The annotation scheme is described in Section 4.)

To measure learning gain over the course of the session, students completed an identical pretest and posttest for each lesson. The average pretest score was 50.98% (min = 23.53%, max = 100%), and the average posttest score was 76.67%

TUTOR Let's move on. [D]

Tutor advances to next task.

TUTOR We have plenty of time. [R]

STUDENT Okay. [ACK]

Student edits code.

STUDENT Which do I put first? [QD]

TUTOR Try it. [D]

TUTOR Be sure you are satisfying the task. [D]

Student compiles, with errors.

TUTOR What you had was close. [FOE]

STUDENT Why did I need quotes for the Hello World println(), but not this one? [QI]

TUTOR Hello World was printing literal "hello world". [AWH]

TUTOR The second was printing the value inside the variable DylansCompGame. [AWH]

STUDENT Oh, alright. [ACK]

STUDENT Makes sense. [FU]

Fig. 2. Sample annotated excerpts from the Lesson 1 corpus

(min = 41.18%, max = 100%), administered immediately after completing the lesson. This learning gain (*posttest* − *pretest*) was statistically significant ($p < 0.0001$). In addition to the pretest, the students also completed a self-efficacy survey with six Likert-scale items prior to the initial tutorial session [17]. Each student's computer science self-efficacy was computed as the average of these six items. The mean self-efficacy score among the students was 3.39 out of a possible 5.00 (min = 2.33, max = 4.33), and as described later, this score is used in the current analysis along with pretest score as control variables within the predictive models of learning.

4 Dialogue Act Annotation

The new refined dialogue act annotation protocol expanded upon a prior scheme for task-oriented tutorial dialogue [17] and was further inspired by previous annotation schemes for tutorial dialogue in several domains [16, 18, 19]. The annotation scheme is presented in detail in Tables 1 and 2, along with the relative frequency of the individual tags and the Cohen's kappa achieved between two independent human annotators. Table 1 displays dialogue acts assigned to both tutor and student utterances; Table 2 displays those assigned to only tutor or only student utterances.

The present dialogue act annotation scheme expands upon a prior set, with a primary goal of further defining the vague large classes previously observed. Figure 3 displays the decomposition of the prior tagset into the current one. The previous set contained thirteen dialogue act tags, with the largest tag accounting for 33.5% of the corpus [17]. The refined annotation scheme presented here contains 31 dialogue act tags, with the largest tag accounting for 13.66% of the corpus. Despite the increased complexity of the proposed annotation scheme, two independent human annotators achieved a Cohen's kappa of $\kappa = 0.87$ on 37% of the corpus (agreement of 89.6%). The prior simpler annotation scheme yielded a Cohen's kappa of $\kappa = 0.79$ (agreement of 81.1%). Of the 31 tags, 21

Table 1. Dialogue act tags assigned to both tutor and student

Tag	Example	Freq.	κ
Explanation (**E**)	*Your code stops on line 2.*	13.66%	0.716
Greeting (**GRE**)	*Have a good day!*	3.49%	0.931
Acknowledge (**ACK**)	*Okay.*	6.93%	0.960
Correction (**CO**)	**explanation*	0.61%	0.734
Observation (**O**)	*See, we have an error.*	1.87%	0.582
Extra Domain Question (**QEX**)	*How are you today?*	1.11%	1.000
Extra Domain Answer (**AEX**)	*I'm doing well.*	1.11%	0.916
Extra Domain Other (**OEX**)	*Calculus is difficult.*	3.59%	0.797
Yes/No Answer (**AYN**)	*No, sir.*	4.20%	0.973
WH-Question Answer (**AWH**)	*Line 9.*	2.68%	0.816

tags achieved a kappa that is characterized as 'almost perfect' inter-rater reliability [20], and the excellent overall kappa achieved by the new tag set suggests that it reliably captures important differences in dialogue acts within the tutorial dialogue corpus. In addition to the tutorial dialogue acts, student task actions were annotated automatically using an edit distance approach. Each period of student coding was classified as improved, worsened, or unchanged, depending on the change in edit distance [17].

5 Relationships between Dialogue and Student Learning

The objective of the present analysis is to identify tutor dialogue act choices correlated with student learning gain. Dialogue acts were identified at the unigram (individual dialogue acts) and bigram (pairs of adjacent dialogue acts) levels [16]. Bigrams were extracted using a three-act collocational window, as demonstrated in Figure 4.

Utterances annotated with the Correction (CO) tag were removed prior to analysis, as these utterances constitute artifacts of the 'instant-messaging' nature of the corpus and reflect typing skill rather than tutoring content. Then, the relative frequencies of each dialogue act tag or bigram were computed, and simple linear correlations were calculated between these and student learning. Then, any correlations that appeared statistically significant at the $p < 0.05$ level were provided as input to a stepwise linear regression model within the SAS statistical modeling software, alongside the pretest and self-efficacy scores. Providing the pretest and self-efficacy as predictors allows the model to account for any differences in posttest scores explainable by these variables.

Several individual dialogue acts or bigrams were significantly predictive of student learning gain. These predictors and their regression coefficients, along with associated p-values within the stepwise linear regression, are listed in Table 3.

Table 2. Dialogue act tags only assigned to one role

Tag	Example	Freq.	κ
	TUTOR		
DIRECTIVE (**D**)	*Test your program.*	9.26%	0.960
INFORMATION (**I**)	*Variable names must be one word.*	7.64%	0.734
REASSURANCE (**R**)	*We have plenty of time left.*	1.01%	0.748
READY QUESTION (**QR**)	*Ready to move on?*	8.65%	1.000
QUESTIONS (**QQ**)	*Any questions?*	1.32%	0.972
FACTUAL QUESTION (**QF**)	*What line is it waiting on?*	1.21%	0.831
OPEN QUESTION (**QO**)	*How can you fix it?*	0.66%	1.000
EVALUATIVE QUESTION (**QE**)	*Does that make sense?*	0.76%	0.933
PROBING QUESTION (**QP**)	*Do you think that looks correct?*	0.40%	0.712
POSITIVE FEEDBACK (**FP**)	*Very good!*	10.72%	0.948
POSITIVE FEEDBACK (WITH ELABORATION) (**FPE**)	*That's a very good approach.*	1.97%	0.729
NEGATIVE FEEDBACK (**FN**)	*No, that's incorrect.*	0.05%	1.000
NEGATIVE FEEDBACK (WITH ELABORATION) (**FNE**)	*That's not the right syntax.*	0.25%	1.000
OTHER FEEDBACK (**FO**)	*That's an okay implementation.*	0.25%	0.800
OTHER FEEDBACK (WITH ELABORATION) (**FOE**)	*That's alright, but you need to fix line 9.*	0.61%	0.952
	STUDENT		
INFORMATION QUESTION (**QI**)	*Why does that happen?*	1.77%	0.917
CONFIRMATION QUESTION (**QC**)	*It's line 6, right?*	2.18%	0.895
DIRECTION QUESTION (**QD**)	*What do I do next?*	1.32%	1.000
READY ANSWER (**AR**)	*Yes, I'm ready.*	8.14%	0.952
UNDERSTANDING (**FU**)	*Oh, that makes sense!*	1.87%	0.847
NOT UNDERSTANDING (**FNU**)	*I don't know why that works...*	0.71%	0.665

The unigram occurrence of tutor directives were negatively correlated with learning gain, aslseen in previous studies [16, 17]. Interestingly, this was the only unigram significantly correlated with learning gain at the $p < 0.05$ level. Two other previously-identified tutorial decisions also emerged as significant: consecutive tutor directives, as seen in a previous study on the same corpus [17] and a tutor information move following a student answer, as seen in the ITSPOKE dialogue corpus [16].

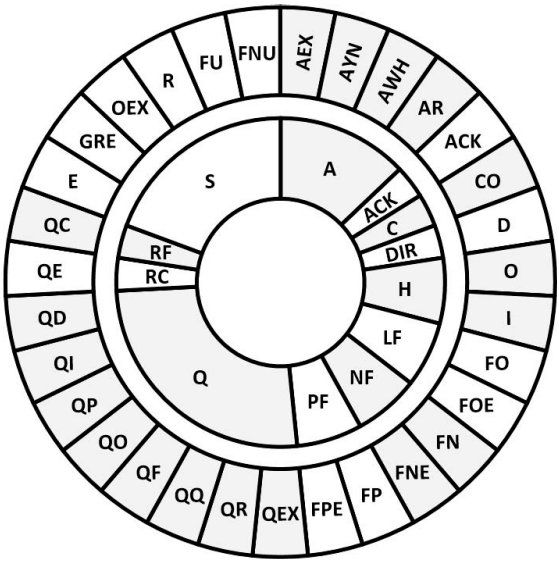

Fig. 3. Decomposition of the prior tags (inner ring) into the new tags (outer ring)

Table 3. A selection of tutor dialogue act choices significantly predictive of student learning gain

Weight	Dialogue Act and Task Sequence	Partial R^2	p
+0.3345	pretest	0.1785	< 0.0001
+0.0868	self-efficacy	0.0003	0.2156
−0.4995	(Improved Code → D (Tutor))	0.0047	0.0050
−0.5004	(FNU (Student) → E (Tutor))	0.0369	0.0049
+0.4388	(QC (Student) → PF (Tutor))	0.0265	0.0153
−0.5781	(D (Tutor))	0.2587	0.0008
−0.5758	(D (Tutor) → D (Tutor))	0.0216	0.0009
−0.4748	(E (Tutor) → QE (Tutor))	0.0231	0.0080
−0.3904	(I (Tutor) → O (Tutor))	0.1474	0.0329
+0.3784	(AWH (Student) → I (Tutor))	0.0907	0.0392
+0.1458	(Intercept)		0.0212

There were several dialogue bigrams significantly correlated with learning gain that had not been identified with a coarser annotation scheme. The bigrams, as shown in Table 3, include improved code followed by tutor directive (D), a student expression of not understanding (FNU) to a tutor explanation (E), a student confirmation question (QC) to positive feedback (PF), a tutor explanation (E) followed by an evaluative question (QE), and a tutor instruction (I) followed by an observation (O). These significant relationships are discussed in the next section.

Role	Utterance	Extracted Bigrams		
Tutor	Do you have any questions? [QQ]			
Tutor	Look over your program. [D]	QQ → D		
Student	No [AYN]	D → AYN	QQ → AYN	
Student	I believe I am understanding the concept. [FU]	AYN → FU	D → FU	QQ → FU

Fig. 4. An example of the collocational window employed to capture dialogue act bigrams at a distance

6 Discussion

This section examines the tutorial dialogue events that were found to be significantly associated with student learning. First we examine tutor directives (D (Tutor)), which are indications that the tutor is giving explicit direction to the student. Consecutive instructions of this nature (D (Tutor) → D (Tutor)) could indicate that the tutor is choosing to exert substantial control over the tutorial session, or that the student is relying heavily on tutor instructions [16, 17]. Another relationship that has been observed in other literature relates to the bigram of a tutor offering information (I) following a student response to a question (AWH), which could indicate a tutor elaborating upon the student's response beyond what he or she initially understood to be correct. This can sometimes provide the answer that the tutor originally expected of the student. This bigram has been previously identified as significant to student learning gain in tutoring for physics [16].

One interesting relationship occurs when the tutor decides to offer a directive after the student has improved the Java program (Improved Code → D (Tutor)). This tutor dialogue act is negatively predictive of learning gain. This relationship could be due to a tutor incorrectly believing that the student needs guidance, and taking control of the session before it is necessary. The directives in the current corpus were frequently an instruction to compile or run the program. This could also occur due the enforced time limit on the session; if the tutor does not believe that the student will complete the lesson before the end of the session, he may give more direct instructions to hasten the completion of the tasks.

A tutor observation after a tutor information turn (I (Tutor) → O (Tutor)) is also negatively predictive of learning gain. This bigram could potentially indicate a "lecturing" mode by the tutor, whereas leaving these tasks to the student to discover could be beneficial to her overall understanding.

Another negative association with learning emerges when tutor evaluative questions, such as *"Does that make sense?"*, follow an explanation (E (Tutor) → QE (Tutor)). One suggested interpretation of this phenomenon is that new students lack meta-cognition; that is, students may not truly know if the material 'makes sense' yet. This is possibly a novice tutor move, as experienced tutors

tend to ask more open-ended questions, judging a student's understanding by his or her demonstrated ability to use the material, rather than relying on the student's meta-cognitive abilities.

Another bigram that was negatively correlated with student learning gain was a tutor offering an explanation when the student expresses a lack of understanding (FNU (Student) → E (Tutor)). This could be explained by a tutor instinctively offering the solution to the student, instead of allowing an exploratory approach by the student before giving aid.

The only bigram found to be significantly positively correlated with learning gain was positive feedback after a confirmation question from the student (QC (Student) → PF (Tutor)). Often, interchanges with these annotations were of the form *"I think the answer is X?"*, followed by a *"Yes, very good!"*. The decision to actively support a student's uncertain answer may provide the student some level of confidence in his ability, which can positively impact further work in the session.

7 Conclusion and Future Work

Tutorial dialogue is rich and highly effective, yet the mechanisms responsible for its effectiveness are not fully understood. Identifying tutor dialogue acts that are associated with student learning gain is a promising direction for research. This paper has presented a novel dialogue act annotation scheme designed to substantially reduce the dominance of a vague majority class that has existed in many prior annotation schemes. When applied in a regression analysis to predict student learning, this new annotation scheme demonstrated its use in identifying previously undiscovered specific dialogue interactions that are predictive of outcomes.

Compelling directions for future work include identifying and comparing effective tutor choices across differing student types, e.g. low versus high self-efficacy students, or students entering from a variety of disciplines. Additionally, a crucial direction for the field is to examine how our annotation schemes support machine learning and data mining on corpora of tutorial dialogue in ways that can inform the design of or support the direct extraction of effective tutorial dialogue system behaviors. These lines of investigation will lead to greater understanding of student learning through tutoring and will inform the design of tutorial dialogue systems.

Acknowledgements. The authors wish to thank Joseph Wiggins for his contributions to the annotation stage of this study, and the members of the Center for Educational Informatics at North Carolina State University for their helpful input. This work is supported in part by the Department of Computer Science at North Carolina State University and the National Science Foundation through Grant DRL-1007962 and the STARS Alliance, CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

1. VanLehn, K., et al.: When Are Tutorial Dialogues More Effective Than Reading? Cog. Sci. 31(1), 3–62 (2007)
2. Bloom, B.S.: The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educ. Res. 13(6), 4–16 (1984)
3. Chi, M.T., et al.: Learning from human tutoring. Cog. Sci. 25(4), 471–533 (2001)
4. Lepper, M.R., et al.: Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. Computers as Cognitive Tools 1993, 75–105 (1999)
5. Graesser, A.C., et al.: Collaborative dialogue patterns in naturalistic one-to-one tutoring. Applied Cog. Psy. 9(6), 495–522 (1995)
6. Chi, M., VanLehn, K., Litman, D.: Do Micro-Level Tutorial Decisions Matter: Applying Reinforcement Learning to Induce Pedagogical Tutorial Tactics. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 224–234. Springer, Heidelberg (2010)
7. Dzikovska, M.O., Steinhauser, N.B., Moore, J.D., Campbell, G.E., Harrison, K.M., Taylor, L.S.: Content, social, and metacognitive statements: An empirical study comparing human-human and human-computer tutorial dialogue. In: Wolpers, M., Kirschner, P.A., Scheffel, M., Lindstaedt, S., Dimitrova, V. (eds.) EC-TEL 2010. LNCS, vol. 6383, pp. 93–108. Springer, Heidelberg (2010)
8. Kumar, R., Ai, H., Beuth, J.L., Rosé, C.P.: Socially Capable Conversational Tutors Can Be Effective in Collaborative Learning Situations. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 156–164. Springer, Heidelberg (2010)
9. D'Mello, S.K., et al.: A Motivationally Supportive Affect-Sensitive AutoTutor. New Perspectives on Affect and Learning Tech. 3, 113–126 (2011)
10. Chen, L., et al.: Exploring Effective Dialogue Act Sequences in One-on-one Computer Science Tutoring Dialogues. In: Tetreault, J., et al. (eds.) Proc. 6th BEA Work., Portland, USA, pp. 65–75. Assoc. for Comp. Ling (2011)
11. Stellan, Ohlsson, o.: Beyond the Code-and-count Analysis of Tutoring Dialogues. In: R, Luckin, o. (eds.) Proc. 13th Int. Conf. AIED, Los Angeles, USA, vol. 158, pp. 349–356. IOS (2007)
12. Forbes-Riley, K., Litman, D.J.: Adapting to Student Uncertainty Improves Tutoring Dialogues. In: Vania, Dimitrova, o. (eds.) Proc. 14th Int. Conf. AIED, Brighton, United Kingdom, pp. 33–40. IOS (2009)
13. Cohen, P.A., et al.: Educational Outcomes of Tutoring: A Meta-analysis of Findings. Am. Educ. Res. J. 19(2), 237–248 (1982)
14. D'Mello, S.K., et al.: Mining Collaborative Patterns in Tutorial Dialogues. J. EDM 2(1), 1–37 (2010)
15. Chu-Carroll, J.: A Statistical Model for Discourse Act Recognition in Dialogue Interactions. In: Chu-Carroll, J., Green, N. (eds.) AAAI Spring Symp.: Applying Machine Learning to Discourse Processing, Pan Alto, USA, vol. 1996, pp. 12–17. AAAI Press (1998)
16. Litman, D.J., Forbes-Riley, K.: Correlations between dialogue acts and learning in spoken tutoring dialogues. Nat. Lang. Eng. 12(2), 161–176 (2006)
17. Mitchell, C.M., et al.: Recognizing Effective and Student-Adaptive Tutor Moves in Task-Oriented Tutorial Dialogue. In: Youngblood, M.G., McCarthy, P.M. (eds.) Proc. 25th Int. FLAIRS Conf., Marco Island, Florida, pp. 450–455. AAAI Press (2009)

18. Person, N.K., et al.: The Dialog Advancer Network: A Conversation Manager for AutoTutor. In: Gauthier, G., et al. (eds.) Proc. ITS Work. Modeling Human Teaching Tactics and Strategies, Montreal, Canada, pp. 86–92. Springer (2000)
19. Core, M.G., Allen, J.F.: Coding Dialogs with the DAMSL Annotation Scheme. In: Proc. 1997 AAAI Fall Symp.: Communicative Action in Humans and Machines, Providence, USA, pp. 28–35. AAAI (1997)
20. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data Data for Categorical of Observer Agreement The Measurement. Biometrics 33(1), 159–174 (1977)