AC 2011-2774: FIRST IMPRESSIONS: THE FIRST TWO POSTS AND THEIR INFLUENCE ON THE DEVELOPMENT OF ONLINE QUESTION-ANSWER DISCUSSION THREADS

Michael Hergenrader, Information Sciences Institute at the University of Southern California

I am currently a junior at the University of Southern California majoring in Computer Science and Spanish.

Joanna Drummond, University of Pittsburgh
Jihie Kim, University of Southern California

Jihie Kim is the Principal Investigator of the Intelligent Technologies for Teaching and Learning group in the USC Information Sciences Institute (http://ai.isi.edu/pedtek). She is also a Research Assistant Professor in the Computer Science Department at the University of Southern California (USC). Dr. Kim received a Ph.D. from the USC, and a master's and a bachelor's degrees from the Seoul National University.

Her current interests include pedagogical discourse analysis, human-computer interaction, social network assistance, and assessment of student collaborative online activities. She leads synergistic work among machine learning experts, educational psychologists, NLP researchers, and STEM instructors. She is the PI of five NSF projects including the CCLI/PedDiscourse, CCLI/PedWiki and NSDL/SocRecomm projects under the EHR Directorate and CreativeIT/PedGames and IIS/PedWorkflow projects under the CISE Directorate. Under the retired PedDiscourse effort, her team designed, deployed and evaluated software tools to assist online dialogue in the context of a discussion board.

First Impressions: The First Two Posts and their Influence on the Development of Online Question-Answer Discussion Threads

1. Introduction

With universities nationwide challenged to provide funding for increasing engineering course enrollment, it seems natural that online courses are becoming more popular[1], cutting costs while still providing students with a college-level education. The switch to these distance learning environments provides not only financial gain for the schools, but also intellectual gain to those without the means to attend a university. Yet, while technically simple to facilitate, this transfer becomes much more complex upon observation of what is lost: face to face connection between students and instructors. In a situation lacking this contact and set meeting times, the interaction changes greatly.

In this paper, we hope to contribute to the study of these new online interactions in the context of online discussion boards, tools that allow students and professors to ask and answer questions in an asynchronous manner. Our study focuses on the first two posts of the discussions – the first usually being a question or issue and the second some response – and their impact on the overall development of the thread. This development provides insight into the degree of student participation and could ultimately hint at the effectiveness of these new tools with regards to student learning.

Drummond and Kim[2] analyzed the role of answers in generating more participation by students. Our work extends their analysis by looking at various question-answer patterns, focusing, in our case, on the beginning of discussions.

We start our analysis by examining the commonly held view that the first post by itself determines how developed a thread will ultimately become. In this context, we define "more developed" as a thread that contains multiple posts and involves multiple users, one that promotes more student participation. Our findings based on the thread data were quite unexpected.

Expanding our analysis to the first two posts, we next observe how different combinations between the two impact the rest of the thread, if more exists. At first, we examine the pair in a general manner, observing various dialogue patterns between them, and then proceed to analyze the specific language patterns that users employ, seeking the particular speech patterns that are more likely to stir up longer threads.

Finally, we turn away from the "what was said" factor towards the "who said it" factor. The language that members use in the second post may affect the overall thread development, but equally important is the role of the user who made the response. Specifically, we contrast the development of threads with the second post written by a student against those with one written by a professor.

2. Study Context

The frameworks of Communities of Practice and Activity Theory have been adopted by practitioners of distributed learning as both aides to understanding and as models for development[3]. Wegerif, using a Communities of Practice framework to analyze accessibility and participation in an online class, found that students' success or failure in the class depended upon the extent to which students participated online, and whether they belonged or felt like an outsider[4]. The study clearly represented a case for the Sense of Community Theory laid out by McMillan and Chavis, in which members' needs of learning were met through sharing ideas, successes, and failures[5].

Golub added to this theory, elaborating that the mutual exploration, meaning-making, and feedback often lead to better understanding on the part of students, and to the creation of new understandings for them[6]. His and Wegerif's studies clearly indicated students needed to become more involved in online message boards via more posts – either posing questions or providing possible answers, correct or not. Suler summed it up best: the more posts students make, the less passive their learning becomes[7]. In short, "lurkers" must become "posters."

Masters and Oberprieler also argued in favor of the "more is better"[8] approach, showing that longer discussions were more likely to provide widespread academic benefit, as the number of students posting would increase[9].With the notion of widespread achievement in mind, we hope to provide factors that can help predict the total length of threads, and by doing so, provide incentive to further analyze how instructors can help bring about longer threads with more active posters, and thus, as advocated by various studies, increase student success.

Our data comes from an operating systems course at the University of Southern California that offers enrollment for both graduates and undergraduates. The intended benefit of the online forums was for students to have questions and issues addressed while others could reference these threads to avoid similar problems. Four team projects dictate the organization of the forums, each containing threads with a "Q&A Style" of discussion, linked up by posts with reply-to relationships. The data consists of 412 of these threads coming from two semesters of the class.

Each of these threads is annotated with Speech Acts, tags that mark general language patterns and their roles in dialogue[10]. For our purposes, we chose to group them into categories of questions, issues, and answers, each containing individual tags that indicate defined text sequences. Marking certain patterns was not always a clear-cut process, as some could be used in a variety of contexts, each with its own meaning. One example of this was the phrase "what if." While this phrase is intrinsically hypothetical in all contexts, on certain occasions it contained a conceptual or reasoning question, while in other situations it presented an inquiry about a process.

To account for indefinite mappings between tags and text, we employed multiple human annotators, with the ultimate goal of eliminating biased or uninformed annotations stemming from one individual. The diversity amongst annotations solved one challenge, but brought another: data consistency.

To solve this potential pitfall, we compared the results of various annotations via Kappa scores, a statistical measure of agreement between two selected annotators that corrects for pure chance[11]. These scores are based on two types of agreement: observed and expected.

The observed part is measured by an in-depth comparison between every post of two data sets. If both annotators marked a phrase in a certain post with the same tag, the level of agreement would increase. The same would occur if neither marked a phrase with any tag. On the other hand, if one annotator marked a speech pattern and the other did not or marked it with a different tag, the level would decrease.

The expected part considers the data sets entirely independent, representing the probability of agreement between the annotators that is based on chance, as if their tagged elements were marked at random. The two parts are represented in this formula[12]:

$$K = \frac{p_{\text{observed}} - p_{\text{expected}}}{1 - p_{\text{expected}}}$$

With 1.0 as a maximum, Kappa scores of approximately 0.70 and above indicate good agreement[13]. As seen in Table 1, the results reflected quite good consistency between the two semester data sets. Scores at the category level are similarly high, indicating agreement if a phrase was or was not tagged with any contained speech act by both annotators. Also included are the speech act tag descriptions and sample phrases seen in posts.

Speech Act Tag	Description	Sample Cue Phrases	Kappa
Question Category	*Represents a question*	N/A	0.94
QCONF	A question asking whether some assumption is correct, or seeking permission	"Right?" "Do we have to", "Is it necessary", "Can I"	0.92
QWHAT	A question about concepts, definitions, facts, or reasoning	"What is", "Why", "When", "Where"	0.82
QHOW	A question about a process	"How to/can/does", "Any ideas?" "What can I do?"	0.83

Q_GENCUE	A generic question or inquiry; usually occurs when a question does not fit the above criteria	"I am wondering", "My question remains", "My query is about"	1.00
Issue Category	*Represents a problem the user is experiencing*	N/A	0.88
DOESNT_WORK	A specific problem that users are experiencing	"error", "fault", "gets stuck", "exact problem", "issue"	0.83
Answer Category	*Represents an answer*	N/A	0.79
INST	Represents a confident suggestion or answer to a question	"You need to", "You should", "What I would do", "This involves"	0.83

Table 1: Relevant Speech Acts annotated in data sets

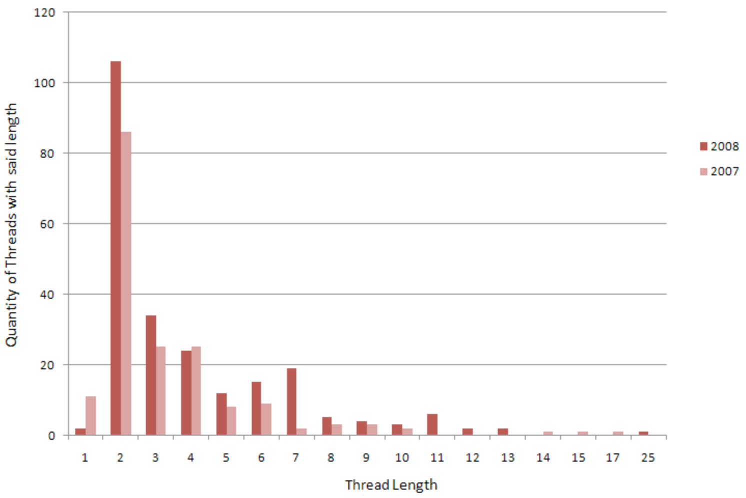

Figure 1: Thread Length Distribution

The purpose of this particular study of thread length arose from an observation consistent across the two examined semesters: a large number of threads with a length of only two posts. The amount of threads with two posts, as seen in Figure 1, is nearly equal to the *total* number of threads with more. We decided to investigate this phenomenon, setting aside the thirteen threads with only one post (analyzing the other 399 of them), labeling the threads with two posts as "less developed" and those with more as "more developed." Noticing this divide, the inquiry into what factors about the initial two posts would influence the thread's development became the logical next step.

3. The Initial Post

To draw conclusions about this, we first hypothesize in accordance with the intuitive claim that the initial questions or issues (the ones belonging to the first post) affect the post count of a thread. To do so, we divide our inquiry into three test categories based on the initial post: the first describing those containing only one type of question, the second containing those with pairs of questions, and the third containing those that express combinations of questions and issue statements, detailing problems a user experiences with his project.

We further test these categories using a two-tailed student's t-test. This test examines if two distributions are different enough that it is unlikely that chance alone would cause this difference. Given a p-value less than 0.05, the distributions are significantly different. With a p-value greater than 0.05, but less than 0.10, these distributions are different at the trend level. Thus, in our case, we examine if the distribution of types of initial posts is significantly different between short posts (i.e. thread length = 2) and long posts (thread length > 2).

In the first category, just three different types of questions appeared on their own in the initial post. The frequencies of the three question types - confirmation questions, "how to" questions, and "what" questions - are enumerated in the following table.

Question Type	Thread length = 2	Thread length > 2
Confirmation	55	54
Process to be taken ("how to" questions)	18*	8
Concept/Reasoning ("what" questions)	14	15

Table 2: Category 1: Speech Act Frequency in initial post (denotes $p < 0.05$)*

Table 2 lists the total number of threads between the two semesters that contained these questions alone and no issues. The only question type with a substantial difference is the "how to" question, leaning in favor of less developed threads. This expresses a result about the particular question type, yet doesn't explain much about threads in general, as this result is only about one question type in one particular category – it is isolated.

In the second category, pairs of question types in initial posts without issues were enumerated. In order to focus on combinations more likely to affect the thread length, we limited our results to pairs of categories that were found more than once. Even so, as was seen in the first category, the results are not substantial enough to suggest that combinations of questions in the first post lengthened the overall discussion. The most commonly occurring pairs are listed in Table 3, along with the total number of threads with specified length.

Question-Question Combination	Thread length = 2	Thread length > 2
Concept/Reasoning - Confirmation	13	17
Process - Confirmation	11	8
General - Process	2	3

Table 3: Category 2: Frequency of combinations of Question Acts in initial post

While the results are revealing in regards to what questions students combine in an initial post, they do not give a meaningful indication as to their affect on the overall thread length.

Finally, initial posts that contained combinations of issues and questions were examined, employing the same threshold used in the previous test. The results are listed in Table 4 below. Though the comparative ratios of the last two combinations may suggest that these patterns lead to less developed discussion threads, the small quantity of observed combinations between two different categories of Speech Act annotations suggests that more factors are required to determine this correlation.

Issue – Question Combination	Thread length = 2	Thread length > 2
Doesn't Work - Concept/Reasoning	11	12
Doesn't Work - Process	12	6
Doesn't Work - Confirmation	6	3

Table 4: Category 3: Frequency of combinations of issues and questions in the initial post

Thus, our hypothesis of the initial post question types influencing the overall thread length cannot be confirmed. In all three tests, the scarcity of one or two-member patterns in the initial post and the small differences between the two levels of development prove inconsequential. The data collected about initial posts suggests that using this factor as a method to determine thread development would be an oversimplification.

4. The Posts Combined

Thus, we turn our attention to the effects that the two posts have when observed together. We hypothesize that the combinations of questions or issues from the first post and answers or clarifications from the response dictate how developed a thread becomes. In fact, we first divide the test into four general Q&A patterns involving questions and issues raised in the first post and whether the second post is primarily a definitive answer or request for clarification. Half of the sixteen possible patterns from these four factors are eliminated because all threads contained at least one question (no initial post contained just issue tags). Furthermore, the pattern set is reduced by very few response posts containing both answers and clarification requests – we regard these as outliers. With the remaining patterns, we focus on how the response type (straight answer against request for clarification) functions with the initial post to affect thread development. These patterns are outlined in Table 5.

Initial Post	Response Post Objective	Thread length = 2	Thread length > 2
Questions, no issues	Answer	93*	77
Questions, issues	Answer	57*	39
Questions, no issues	Clarification request	3	8
Questions, issues	Clarification request	1	9*

Table 5: Frequency of threads with general patterns contained in first two posts (denotes $p < 0.05$)*

The data in this table evidences the clear trends in thread development when the response post objective changes, supported further by the significant p-values gathered from the same two-tailed student's t-test that was conducted for the first hypothesis. For responses with a straight answer and no clarification questions, the threads tended to less developed, whereas threads with clarification requests caused threads to expand beyond two posts. Both of these statistical results are consistent with qualitative observation. If an answer to initial questions or issues is given, it becomes less likely that other members will follow up on the thread, doing so mainly in cases of correction or expressed confusion. On the other hand, clarification questions will usually warrant more responses in an attempt to solve the problem at hand.

At this stage, the hypothesis seems to be confirmed, but we delve into specific Speech Act combinations that represent these patterns to solidify it.

For the first pattern, we filter out all initial posts with issue annotations and focus solely on the question-answer relationship. The most common Speech Act pattern turned out to be QHOW-INST, a sensible pair for Q&A project forums, given that it represents an inquiry of a process and a definitive suggestion about what to do. A total of thirteen instances were found for threads of length two, compared to only six for threads of more development. This language pattern helps to explain the general one given in Table 5.

The second pattern allowed all initial posts to contain both questions and issues while the response contained only straightforward answers. Once again QHOW-INST proved to be the dominant pattern in addition to QWHAT-INST.

Question-Answer Combination	Thread length = 2	Thread length > 2
Process – Instruction (QHOW-INST)	21	13
Concept/Reasoning – Instruction (QWHAT-INST)	18	13

Table 6: Frequency of threads with Question-Answer pairs in first two posts (initial posts contain issues)

As Table 6 shows, a response containing confident answers to non-confirmation questions appears to promote the trend of less developed threads, indicating that the first response to a question or issue may influence the thread length more than the initial question itself. Qualitatively speaking, perhaps users who see these questions answered with this authority are inclined to believe the second post is sufficient, and thus do not provide more input or questions. The risk of this case, as Drummond and Kim examined, is that these answers “may not provide enough information to the information seeker.”[14]

The latter two general patterns reflected the opposite trend of the first two – clarification questions extended the length of the thread. To measure this, we again observed Speech Act pairs – the first one being the dominant question of the initial post, and the second the most common question of the response.

In the third pattern, we once again filtered out question posts that contained issue annotations, focusing on more conceptual questions expressed by QWHAT in the initial post. In the response posts, we found that QWHAT and QCONF (clarification questions) were the most common responses to these questions, highlighting the possibility that the first question was not phrased with the utmost clarity or precision, warranting thus a longer thread. Though the amounts of these pairs expressed in Table 7 are not large, they still follow the general trend of longer threads reflected in the third pattern.

Initial Question Type - Clarification Question Type	Thread length = 2	Thread length > 2
Concept/Reasoning – Concept/Reasoning	3	5
Concept/Reasoning – Confirmation	1	3

Table 7: Frequency of threads with Question-Clarification Question pairs in first two posts

This trend became more amplified in the results from the fourth pattern, representing initial posts that raised issues with accompanying questions. Concept or reasoning questions again resulted as the most frequent clarification question, appearing in six threads that were greater than length two, and in *none* that were less developed.

Our hypothesis is thus confirmed based on the general and specific results that various combinations of the first two posts helped to determine. This result seems to give credence to the conclusions reached about the effect of dialogue by VanLehn, Graesser, Jackson, Jordan, Olney, and Rosé: that a dialogue, a discussion containing questions in *both* posts instead of just questions and immediate answers, is more likely to increase active participation among students[15], and in our case, improve the development of a thread. Thus, the combination of the initial post and initial response post is a good indicator of how long the discussion length will become.

5. Answers and Authorship

An even more accurate determinant of length was the identity of the user who posted the first response. We hypothesized that a difference existed between threads in which students responded first and those in which professors responded first. The language patterns between the students and professors were kept constant to make this test - student responses were only tabulated if they used Speech Acts the professor regularly used, mainly INST.

With this constant, it became obvious that the instructor's early responses clearly had an effect on the development of a thread, shortening its length. Of the 399 total threads, the professor for this course was the first to respond 171 times. 113 of these responses, about 66%, were the last posts of their corresponding threads. For those in which a student responded first, the frequency of more developed threads rose dramatically. When students replied first, 25 threads were of length two and 79 of greater length, so only 24% of student initial responses were also the final posts of the thread.

Our hypothesis of the professor's early presence affecting the thread's development is thus confirmed. The conclusion reached by Mazzolini and Maddison is quite similar: "the more instructors posted to discussion forums, the shorter were the discussion threads on average. Instructors who were active… did not appear to stimulate more discussion, and may actually have limited the amount of discussion and the length of discussion threads."[16]

In addition to the thread length increasing when students were first to post, the number of different students posting in the threads also increased: from 15 to 24, evidenced by the larger amount of different user ID's collected from posts. This change implies an interesting direct relationship between student participation and thread length. In view of this data set, it becomes obvious that online student participation is lacking in this particular case. If the professor were to lessen his authoritative presence, either by posting less at the beginning of a thread or by extending it through dialogue-inducing techniques described by VanLehn, Graesser, Jackson, Jordan, Olney, and Rosé[17], perhaps student participation, and thus the effectiveness of the forums, would be improved.

An analysis similar to this one, conducted by McLaren, Scheuer, De Laat, Hever, De Groot, and Rosé, describes an even more effective role of the professor in electronic discussions: being a forum moderator. Machine generated annotations like Awareness Indicators would assist professors in this role, alerting them of student discussion behavior in multiple threads[18]. With this role, a professor limits his active presence, posting mainly to encourage more critical thinking or to invite other student responses. This paper and the related analyses suggest that in online discussions, a professor's role as a supervisor may be more effective. In this case, it is up to students to take charge of their own learning.

6. Discussion and Future Work

In this paper we found that two of the three presented hypotheses accurately explained how the development of a thread became affected in the context of online discussions. Surprisingly, the question or series of questions and issues presented in the initial post did not provide conclusive evidence that the language of the first post influenced the thread length. The next hypothesis tested was found to be confirmed, with the conclusion that the first response to the initial question may be a significant part to determine the overall development of the thread. And the role of the user that posted this next response could also be responsible for how long the thread would become – the presence of the professor early in a thread more often than not signified that the length would be shorter. In addition, it added the side effect of more sparse student participation, giving credence to the suggestion by Palloff and Pratt: "[in an online context] faculty need to be willing to give up a degree of control and allow the learners to take the lead in the learning activities."[19]

In addition to the instructor abstracting his presence during the initial stage of discussions, Roper found that more "meaningful"[20] messages posted by students encouraged everyone to continue participating in the dialog, thus extending its length. These posts went beyond simple compliments to include well-crafted questions and answers - as expected - and also had students actively seeking answers from their peers rather than the instructor[21]. This result supports the theories laid out by Wegerif, McMillan, Chavis, and Golub: large, active communities in a thread allowed an environment for each individual to comfortably become involved. Here, longer threads correlated to widespread involvement, and thus, a greater resource for students.

Yet while the survey suggested that instructors be absent in some respects, it laid out a more optimal role for them: asking the *initial* questions and intervening with more questions only if students had not already done so initially with their peers[22]. In the future, we would like to investigate this proposal statistically, as we have done in Section 5. Our current data sets for this course had no posts with the instructor as the initial poster, but it would be possible given data sets from engineering courses with instructors with this more active approach.

And to continue this particular analysis we would like to build off the idea of student participation expressed in Section 5, investigating whether better developed discussion threads that actively involve many students contribute to their overall project grade and class performance. We would seek a comparison between those who actively participated in the online boards and those who did not or did so minimally. This study, similar to the one done by Wegerif, would perhaps allow us to evaluate the effectiveness of these tools with respect to student learning.

In addition, we wish to find out through use of the Speech Acts whether all the questions that a user poses in the first post are sufficiently answered in the reply or replies following it, an issue studied by Drummond and Kim[23]. To assist in this question, we can take advantage of Speech Act classifiers developed by Ravi and Kim[24], tools that can automatically analyze discussion data sets. These classifiers can enable us to efficiently process a lot more data via machine learning and thus provide even more representative results. Continuing to explore question-answer patterns with accurate results will ultimately help instructors to better diagnose student needs in a virtual classroom context.

1 Ahem, T.C., Cooper, S., Lan, W., Liu, X., Shaw, S., Tallent-Runnels, M.K., and Thomas, J.A. (2006). Teaching Courses Online: A Review of the Research. *Review of Educational Research*, 76: 1, 93-135.

[2] Drummond, J., Kim, J. (2011). Role of Elaborated Answers on Degrees of Student Participation in an Online Question-Answer Discussion Forum, *American Educational Research Association* (to appear).

[3] Cole, M., and Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations*. New York: Cambridge University Press.

[3] Engeström, Y. (1987). *Learning by Expanding: An Activity Theoretical Approach to Developmental Research*. Helsinki: Orienta-Konsultit Oy.

[3] Lave, J. (1996) Teaching as Learning, in Practice. *Mind Culture and Society*, 3: 3, 9-27.

[3] Lave, J., and Wenger, E. (1991). *Situated Learning. Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

[4] Wegerif, R. (1998). The social dimension of asynchronous learning networks. *Journal of*

Asynchronous Learning Networks. 2: 1, 34-49.

[5] McMillan, D.W., and Chavis, D.M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology*, 14: 1, 6-23.

[6] Golub, J. (Ed). *Focus on Collaborative Learning*. Urbana, IL: National Council of Teachers of English, 1988.

[7] Suler, J. (2004). *CyberPsychology and Behavior*, 7, 397-403.

[8] Masters, K., Oberprieler, G. (2004). Encouraging equitable online participation through curriculum articulation, *Computers & Education,* 42, 319-332.

[9] Masters, K., Oberprieler, G. (2004).

[10] Searle, J., (1969). *Speech Acts*. Cambridge: Cambridge University Press.

[11] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20: 1, 37-46.

[12] Forzano, Lori-Ann B., Gravetter, Frederick J. (2009). *Research Methods for the Behavioral Sciences*. Belmont: Wadsworth Cengage Learning.

[13] Cohen, J. (1960).

[14] Drummond, J., Kim, J. (2011).

[15] VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., and Rosé, C. (2007). When Are Tutorial Dialogues More Effective Than Reading?. *Cognitive Science*, 31: 1, 3-62.

[16] Mazzolini, M., and Maddison, S. (2003). Sage, guide, or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers and Education*, 40: 3, 237-253.

[17] VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., and Rosé, C. (2007).

[18] McLaren, B., Scheuer, O., De Laat, M., Hever, R., De Groot, R., and Rose, C. (2007). Using Machine Learning Techniques to Analyze and Support Mediation of Student E-Discussions. *Proceedings of AIED-2007.*

[19] Palloff, R., and Pratt, K. (2001). *Lessons from the Cyberspace Classroom: The Realities of Online Teaching*. San Francisco: Jossey-Bass.

[20] Roper, A. (2007). How Students Develop Online Learning Skills. Educause Quarterly, 30: 1, 62-65.

[21] Roper, A. (2007).

[22] Roper, A. (2007).

[23] Drummond, J., Kim, J. (2011).

[24] Sujith R., Kim, J. (2007). Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers. AIED 2007: 357-364