Initiative In Tutorial Dialogue

Initiative in Tutorial Dialogue*

Mark G. Core and Johanna D. Moore and Claus Zinn

Division of Informatics,
University of Edinburgh, 2 Buccleuch Place
Edinburgh EH8 9LW, UK
[markc|jmoore|zinn]@cogsci.ed.ac.uk

Abstract

One-on-one human tutoring has been shown to be a very effective method of instruction. Many researchers have argued that good human tutors encourage knowledge construction by getting students to do as much of the work as possible and to maintain a feeling of control, while providing students with enough guidance to keep them from becoming too frustrated or confused. So-called "Socratic tutoring" has been proposed as the best way to accomplish these tasks. To implement Socratic tutoring in a dialogue-based ITS we must understand it in terms of the features that characterize dialogue systems such as: dialogue strategies, initiative management, dialogue acts, and turn taking. In this work we investigate initiative in tutorial dialogue by comparing tutoring sessions with a Socratic style to sessions with a didactic tutoring style. Results show that the Socratic dialogues were more interactive than the didactic dialogues; on average students spoke more in the Socratic condition, and tutors asked more questions and made fewer statements. Students learned more in the Socratic condition but due to small student numbers this can only be regarded as a trend. Surprisingly, students took more initiative in the didactic dialogues; unlike we expected, student initiative was not the key element that makes Socratic tutoring effective.

keywords: initiative annotation, initiative analysis, tutorial dialogue

1 Introduction

Studies show that one-on-one human tutoring is more effective than other modes of instruction. The average student who received one-on-one tutoring with an expert tutor scored 2 standard deviations above the average student who received standard classroom instruction (Bloom, 1984). Current intelligent tutorial systems relying on graphical user interfaces reliably produce effect sizes of 1 standard deviation above students only receiving traditional instruction (Anderson, Corbett, Koedinger, & Pelletier, 1995).

What is it about human tutoring that better facilitates learning? Many researchers argue that it is the collaborative dialogue between student and tutor that promotes the learning (Merrill, Reiser, & Landes, 1992a; Fox, 1993; Graesser, Person, & Magliano, 1995). Through collaborative dialogue, tutors can intervene to ensure that errors are detected and repaired and that students can work around impasses (Merrill, Reiser, Ranney, & Trafton, 1992b). Previous research has also shown that students must be allowed to construct knowledge themselves to learn most effectively (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & Lavancher, 1994; VanLehn, Siler, Murray, & Baggett, 1998). The consensus from these studies is that experienced human tutors maintain a delicate balance allowing students to do as much of the work as possible and to

*The research presented in this paper is supported by Grant #N00014-91-J-1694 from the Office of Naval Research, Cognitive and Neural Sciences Division. Thanks to Jean Carletta and our reviewers for their comments on this work.

maintain a feeling of control, while providing students with enough guidance to keep them from becoming too frustrated or confused.

So-called "Socratic tutoring" has been proposed as the best way to accomplish these tasks. But what are the features that characterize Socratic tutoring dialogues? To implement Socratic tutoring in a dialogue-based ITS we must understand it in terms of the features that characterize dialogue systems such as: dialogue strategies, initiative management, dialogue acts, and turn taking. In this paper, we explore the role of initiative in Socratic dialogue and study the relationship between initiative and learning.

Our experimental paradigm consists of collecting dialogues between students and a Socratic tutor (a tutor who asks questions instead of giving explanations) and contrasting them with dialogues between a student and a didactic tutor (a tutor who gives explanations followed by questions to test student understanding). Our hypotheses were that the Socratic sessions would be more successful and the reason for their success would be that students were encouraged to take more initiative.

2 Previous Work

2.1 Defining Initiative

Sinclair and Coulthard (1975) developed a dialogue grammar for classroom discussions. Their minimal unit of dialogue is the exchange which is composed of an initiating move, an optional responding move, and an optional feedback move. Whoever makes the initiating move is said to have initiative for the exchange. Moves consist of an uninterrupted sequence of words by a speaker. Moves have a structure of their own but we will discuss only the "head" element (primary function) of moves. The head elements of initiating moves can be questions, directives, or informative statements; informative statements are distinguished from replies in that their primary purpose is to give information, not to respond. By primary purpose, we mean the function of the move as mutually agreed upon by the dialogue participants.

Sinclair and Coulthard describe cases of "re-initiation" where the student replies incorrectly and the original initiating question is reasked. When the tutor accepts the student answer then the exchange is over. However, re-initiation does not capture all the ways an exchange can be disrupted. Linell *et al.* (1988) discuss how a responder can ask for clarification, challenge the speaker, and change topics as well as respond directly to an initiating move. Linell *et al.* do not assign initiative directly to speakers but instead rank speaker moves based on how much "they can be regarded as governing or steering the ensuing dialogue and as being governed or commanded by the preceding dialogue" (p. 419). For example, an utterance which is not a response in any way but requires a response from the listener is ranked highest with a value of six. Minimal responses are at the other end of the scale (with a rank of two); they invite no response and give no more information than required.

Shah (1997) defines student initiative more directly as "any contribution by the student that attempts to change the course of the [tutoring] session" (p. 13). Shah's goal was to study cases where student behavior deviated from the tutor's expectations as encoded in the tutor's discourse plan as defined in (Freedman, 1996). In terms of annotating student initiative, Shah takes a practical approach tailored to her corpus of remediation dialogues. In these dialogues, tutors would quiz students about the answers they gave during problem solving (rather than having students self-explain or actually problem solve). In this corpus, student initiatives are student utterances that are not answers to questions. Shah assumes that these initiatives are dealt with exclusively by the next tutor speech act and that the tutor then takes back initiative. Thus, the tutor always has initiative except during student initiatives and the tutor's responses to those initiatives.

Chu-Carroll and Brown (1998) state that it is important to differentiate initiative (they call it dialogue initiative) from task initiative. They define dialogue initiative by stating that it "tracks the lead in determining the current discourse focus" (p. 6)[1] and that task initiative "tracks the lead in the development of the agents' plan" (p. 6). Presumably, determining the discourse focus means something like setting the discourse segment purpose in Grosz and Sidner's (1986) theory of discourse. What it means to take the lead in developing the agents' plan depends on the plan representation but informally can refer to adding or taking away actions from the plan, rearranging actions, or setting parameters.

Although Chu-Carroll and Brown claim that dialogue and task initiative can be annotated directly from these definitions, it is worth noting that several research projects (Strayer & Heeman, 2001; Jordan & Di Eugenio, 1997; Doran, Aberdeen, Damianos, & Hirschman, 2001; Walker & Whittaker, 1990) have adopted the dialogue-initiative annotation guidelines proposed in (Whittaker & Stenton, 1988). Whittaker and Stenton refer to dialogue initiative as control of the dialogue and define rules for determining who has control (*e.g.,* commands mean the speaker has control, questions mean the speaker has control unless the question follows a previous question).

2.2 Studies of initiative in human-human corpora

Work in computational linguistics (Walker & Whittaker, 1990) has successfully used initiative as a tool for dialogue analysis. Walker and Whittaker studied two dialogue genres, advisory dialogues (ADs) and task oriented dialogues (TODs). In the ADs they examined, novices were asking experts for help with either their finance or software problems. In the TODs, an expert was guiding a novice through the construction of a water pump over the telephone or via keyboard (chat). Walker and Whittaker labeled these dialogues for initiative based on the guidelines in (Whittaker & Stenton, 1988).

In the TODs, most of the time the expert has initiative; the ADs have closer to an equal sharing of initiative. In TODs, the expert had initiative 91% of the time while only having initiative 60% of the time in finance ADs and 51% of the time in software ADs. This distribution reflects the fact that in ADs, the user must provide the expert with details of the problem to be solved. The expert would let the novice describe the problem in their own words rather than continually prompting them for the information.

Another issue Walker and Whittaker investigated was how initiative was passed from one speaker to another. Of particular interest are abdications, initiative changes marked by prompts - utterances with no propositional content. They directly signal the listener to take initiative. Abdications were more prevalent in the TODs; novices would interrupt experts to report problems and then use abdications to signal the end of the interruption. The ADs were more collaborative and replied less on abdications.

Shah (1997) investigated initiative in CIRCSIM dialogues, typed human-human tutoring dialogues dealing with the circulatory system. Her corpus consisted of students' initial tutoring session and a subsequent session with each of the same students. She labeled any student utterance that was not an answer to a question as an initiative; these were then categorized based on communicative goal (*e.g.,* challenge, support, repair, request information). Shah found that the initial sessions have twice the number of student initiatives as the set of subsequent sessions. The nature of the student initiatives changes as well. The proportion of student initiatives associated with confusion (long pauses and self repairs) decreases in subsequent sessions and the proportion of challenges increases. Shah also looked at tutor reactions to student initiatives; she found that tutors sometimes rejected student initiatives, but she did not investigate what triggered such actions.

(Graesser & Person, 1994) labeled student questions (a subset of the initiatives studied by Shah) in a corpus of tutoring sessions for a research methods course. Graesser and Person developed a taxonomy of different question types. Of specific interest are deep-reasoning and knowledge deficit questions. Deep-reasoning questions involve causal reasoning and hypothetical situations. Knowledge deficit questions are triggered when a student realizes an inconsistency or gap in his understanding or gets stuck on a problem. Graesser and Person found that in the first half of the

[1] The page numbers come from the digital version: http://citeseer.nj.nec.com/244268.html

course there was a negative correlation between overall number of student questions and exam scores. In the second half of the course, there were positive correlations between exam scores and the proportion of student questions that were deep-reasoning questions and the proportion of student questions that were knowledge deficit questions.

Our study focused solely on initiative and did not address the difficult problem of categorizing question semantics. Initiative is a noisy measure of student participation. Shallow questions such as "What do I do next?" are treated the same as insightful questions such as "Is a load basically the opposite of a source?". Despite this interference, we hoped that high levels of initiative would characterize students who took control of their learning and as a result scored well in the post experiment test.

3 Study of Socratic v. Didactic Tutoring

In our study we created a corpus of Socratic and didactic tutoring dialogues. The first hypothesis to be tested was that the experimental manipulation was successful - that Socratic dialogues were more interactive than the didactic dialogues. The second hypothesis was that the Socratic dialogues were more successful in promoting learning. The third hypothesis was that more student initiative is a critical part of what makes Socratic tutoring successful; although if the student takes too much initiative we would expect performance to drop off based on studies indicating the problems with unstructured exploratory learning (Pea & Kurland, 1984; Pirolli & Anderson, 1985).

3.1 Method

The setting for this study is a course on basic electricity and electronics developed with the VIVIDS authoring tool (Munro, 1994). Students read textbook-style lessons and then perform labs using a circuit simulator with a graphical interface. (Rosé, Moore, VanLehn, & Allbritton, 2000) describes an experiment where students went through these lessons and labs with the guidance of a human tutor. Before the lessons students were given pretests to gauge their initial knowledge. After two tutoring sessions, students took the same tests again. We refer to the difference in their scores as learning gain. There were three sets of tutoring sessions:[2] (1) the trial sessions where the tutor was not given any instructions on how to tutor, (2) the Socratic sessions where the tutor was instructed not to give explanations and instead ask questions, and (3) the didactic sessions where the tutor was encouraged to give explanations and then probe student understanding with questions. During these sessions, the student and tutor communicated through a chat interface. We will refer to the logs of this chat interface as the BEE dialogues. They are publicly available at `http://www.cogsci.ed.ac.uk/~jmoore/tutoring/BEE_corpus.html`.

In section 2.1, we reviewed various definitions of initiative. From these definitions, we chose Chu-Carroll and Brown's (1998) definition of (dialogue) initiative. The alternative would be using Linell *et al.*'s definition of initiative (1988). It is not clear whether Chu-Carroll and Brown's description of initiative as setting the discourse segment purpose is superior to Linell *et al.*'s dialogue move ranking. In the end, we chose Chu-Carroll and Brown's definition because of the availability of easy-to-use annotation guidelines (Whittaker & Stenton, 1988). Note, we do not attempt to annotate task initiative here; in the discussion section we mention future work on annotating scaffolding and question and answer types. This work should allow us to measure task initiative as well.

In the rest of this section we give details of Whittaker and Stenton's annotation guidelines. Whittaker and Stenton define a set of rules for assigning initiative to every turn (uninterrupted sequences of words by one speaker) in a dialogue. The turn must be classified into one of the following types based on its main purpose:

- **assertions** — declarative turns used to state facts.

[2] By session, we mean all the dialogue between the tutor and one particular student.

- **commands** — turns intended to instigate action.
- **questions** — turns intended to elicit information.
- **prompts** — turns not expressing propositional content (*e.g.,* “yeah”, “okay”).

We use the rules below to assign initiative. These are the same as the rules given by Whittaker and Stenton except that we make the assumption that a statement following a question responds to that question. See our corpus web page for more details.

```
if turn = command then speaker has initiative
if turn = question then
   if last_turn = question or command then listener has initiative
                                      else speaker has initiative
if turn = statement then
   if last_turn = question then listener has initiative
                           else speaker has initiative
if turn = prompt then listener has initiative
```

A benefit of this annotation scheme is that in our corpus the majority of turns can be automatically labeled: **question**s often ended in question marks; **command**s often started with verbs; a list of common **prompt**s (“okay”, “yeah”) allowed most of these to be labeled, and **statement** could be used to label everything else.

We needed human annotators to correct the automatic labeling. One of the authors of the paper and another annotator (not a project member) corrected the utterance type annotations (the author only corrected 3 out of the 23 dialogues used in this study).

The annotators had a reference manual and trained on trial sessions of the dialogues. To test interannotator reliability, the author and external annotator labeled the same 757 utterances of non-training data; the resulting interannotator reliability as measured with the kappa statistic was 0.92. Generally, kappa values above 0.8 are considered acceptable and values between 0.8 and 0.67 marginal.[3] To download the annotation manual or get more details of the methodology consult the web page, `http://www.cogsci.ed.ac.uk/~jmoore/tutoring/BEE_corpus.html`

3.2 Analyses and Results

3.2.1 Were the Socratic dialogues really Socratic?

Following the methodology of (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), we verified that the experimental manipulation was successful - that the Socratic dialogues were more interactive. Chi *et al.* look at the pure quantity of language from student and tutor as well as the number of moves such as self-explaining, giving feedback, asking a question, and scaffolding. We also started by looking purely at the quantity of language output by the tutor and student. Our test data consists of 23,451 words, 2853 utterances and 1547 turns of Socratic dialogue and 26,195 words, 2993 utterances, and 1378 turns of didactic dialogue.

We first focused on the number of words spoken, in particular, the percentage of words spoken by the student. The average per session for the didactic dialogues was 26%; the average for the Socratic dialogues was greater, 33% (t=2.26, df=18, $p < 0.05$). Another relevant statistic is the average number of words per tutor utterance. The average per session for the didactic dialogues was 9.42 words/utt and the average for the Socratic dialogues was less, 8.4 words/utt (t = 2.33, df=18, $p < 0.05$).

Following Chi *et al.*, we also look at the average number of utterances per turn; if the Socratic dialogues were more interactive, then we should see fewer tutor utterances per turn. More tutor utterances per turn would suggest long explanations. Indeed, the Socratic dialogues had a lower average tutor utts/turn, 2.38 than the didactic dialogues, 3.02 (t = 4.9, df=18, $p < 0.05$).

[3] These guidelines are based on comments by Krippendorff (1980) (as summarized in (Carletta, 1996)). Krippendorff considered the case of two annotated variables. He said that comparisons were reliable when the kappas for those variables were above 0.8. If the kappas were between 0.6 and 0.8, then tentative conclusions can be drawn.

	Question	Statement	Command	Prompt
Didactic	29%	47%	9%	15%
Socratic	42%	34%	6%	18%

Table 1: Breakdown of Tutor Moves

	Question	Statement	Prompt
Didactic	15%	79%	6%
Socratic	10%	88%	2%

Table 2: Breakdown of Student Moves

Although we do not have the detailed annotations of (Chi et al., 2001), we did annotate for **question**s, **statement**s, **command**s, and **prompt**s. In Table 1, we see that tutors did ask significantly more **question**s in the Socratic condition (t = 5.85, df=18, p < 0.001) and issue significantly fewer **statement**s (t = 4.16, df=18, p < 0.001). The differences in **command**s and **prompt**s are not significant (t = 1.99, df=18, NS and t=1.48, df=18, NS).

Table 2 shows a breakdown of student moves; student commands are not included because only two occurred in the whole corpus. The difference between student questions in the two conditions is not significant (t = 1.56, df=18, NS) nor is the difference between student prompts (t= 1.83, df=18, NS). There is a significant difference between student statements in the two conditions (t=2.27, df=18, p < .05) reflecting the earlier result that students spoke more in the Socratic condition. Perhaps with more data the differences in questions and prompts would be greater. Currently we cannot draw any conclusions from the breakdown of student moves.

Although this analysis is course grained, it will tend to blur the differences between Socratic and didactic dialogues rather than create a false distinction. The fine grained analysis used in (Chi et al., 2001) uses categories such as scaffolding and explanation. Scaffolding includes tutor statements that hint rather than explain. In our analysis all statements are regarded equally. Despite this fact, we do see a difference in the percentage of statements and questions made by the tutor in the two conditions. This evidence along with the results on quantity of language shows that the Socratic dialogues were more interactive.

3.2.2 Are Socratic dialogues more effective?

This analysis is reported on in (Rosé et al., 2000) and summarized below. In terms of raw learning gain data, the Socratic dialogues were more effective; the effect size ([mean Socratic learning gain - mean didactic learning gain]/standard deviation of didactic gain score) was 1 standard deviation in favor of Socratic tutoring. Small student size means this result can only be taken as a trend as verified using an ANCOVA with dialogue type as the independent variable, pre-test score as the covariate, and post-test score as the dependent variable F(1,18) = 3.13; p < .1.

(Chi et al., 2001) also investigated the differences between Socratic and didactic tutoring. They found that students performed equally well in the two conditions. However, a possible confound in favor of didactic tutoring was that tutors when in the didactic condition sometimes gave away post-test question answers in their explanations. Students tutored with the Socratic method had to do more work to score equally well on the post-test.

Thus, the evidence favoring Socratic tutoring is weak and work still needs to be done to verify its effectiveness. In the meantime, we can still analyze didactic and Socratic dialogues in terms of the features that characterize dialogue systems, in particular, initiative management. Eventually we will need to implement one of these two methods in our tutorial dialogue system and will need to know how to manage initiative.

3.2.3 Initiative analysis

Our first analysis was to measure the average percentage of turns for which students had initiative in the Socratic and didactic dialogues. Surprisingly, students had initiative for fewer turns on

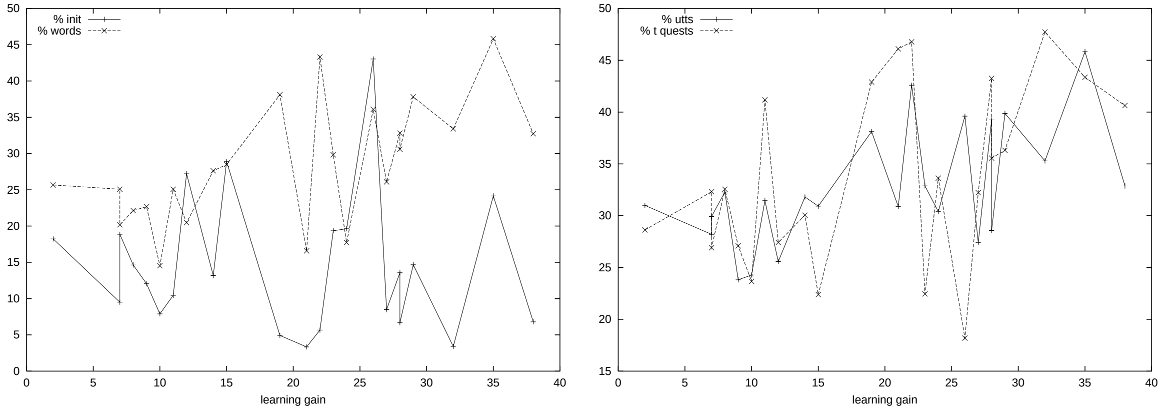

Figure 1: Learning Gain Comparisons

average (10%) in the Socratic dialogues than in the didactic dialogues (21%).[4] Informal analysis of the dialogues showed that students did not take advantage of the fact that Socratic dialogues were more interactive; they did not often take over the conversation. In the didactic dialogues, we noticed more student questions (initiatives) following explanations than in the Socratic dialogues. Perhaps the long explanations confused students.

We next tested the relationship between initiative and learning gain. Since initiative factored more heavily in didactic dialogues we hypothesized that the degree to which students took initiative might correlate negatively with learning gain. Since Socratic and didactic dialogues also differ in interactivity, we tested the relationship between learning gain and the interactivity measures of average percent of words and utterances spoken by the student and average percent of tutor utterances that were **question**s. Figure 1 shows this data; the left graph shows that initiative varies erratically as learning gain increases; there is no relationship (Pearson's r=-.0689, n=23, NS) between these variables. The left graph also shows average percentage of words spoken by the student; it does have a relationship with learning gain (Pearson's $r = 0.6$, $n = 23$, $p < 0.005$). The right graph shows the relationship between percentage of utterances spoken by the student and learning gain (Pearson's $r = 0.56$, $n = 23$, $p < 0.005$), and the relationship between average percentage of tutor utterances that were **question**s and learning gain (Pearson's $r = 0.46$, $n = 23$, $p < 0.05$). Since percentage of tutor utterances that were **statements** did not have a relationship with learning gain (Pearson's r = -.3143, n=23, NS), we did not plot these results. We focused on tutor questions and statements because in section 3.2.1 we found significant differences between the Socratic and didactic dialogues in these utterance types.

In section 2.2, we discussed the work of Walker and Whittaker on investigating initiative in the genres of advisory dialogues (ADs) and task oriented dialogues (TODs). Walker and Whittaker also investigated the difference between TODs in a spoken (telephone) modality and a typed (computer chat) modality. The results of their study are shown in columns 3-6 of Table 3 and the corresponding measures from our study are in columns 1 and 2. The Socratic dialogues have almost the same average expert initiative as TODs. In the TODs, the expert would issue a series of commands. In the Socratic dialogues the tutor was issuing a series of questions.

The second row of the table shows average percentage of initiative changes that were abdications. Abdications are the use of prompts to give away initiative; these often occur after interruptions[5] to signal the original speaker to continue. Walker and Whittaker noted that TODs had more interruptions and thus more abdications in the spoken modality. However, the typed TODs had fewer abdications than either the spoken TODs or the ADs. Modality has an impact on how initiative is managed.

In the didactic and Socratic dialogues (both of which are typed) shown in columns 1 and 2,

[4] To analyze significance, we looked at average percentage of expert initiative per session rather than per corpus. For the didactic dialogues, this average is 82% and for the Socratic dialogues it is greater 90% ($t = 2.26$, df=18, $p < 0.05$ two-tailed.

[5] Walker and Whittaker define interruptions as taking the initiative without invitation. It does not refer to interrupting the utterance of the other speaker.

	Didactic	Socratic	AD Finance	AD Software	TOD Phone	TOD Key
Expert-Initiative	79%	90%	60%	51%	91%	91%
Abdication	2.32%	0.43%	38%	38%	45%	28%

Expert-Initiative - % of total turns with expert initiative
Abdication - % of initiative shifts that are abdications

Table 3: Initiative Measures for six Corpora

we see that abdications are rarely used. A number of reasons are possible. In the typed TODs, communication consisted of two simultaneously updated channels. In the tutoring dialogues, participants would send each other short messages. This modality, typed text and restricted turn taking might have reduced the number of abdications. Another possible factor is that students in this study were relatively passive; the tutor could not rely on them to take initiative if she uttered a prompt. The tutor's initiative management also played a role. In our dialogues, after the student took initiative, the tutor would address the student's turn and then often take back inititive not giving the student a chance to utter a prompt.

4 Discussion

Although we make some simplifying assumptions in annotating initiative, it is unlikely that a more complex annotation scheme will change these results. Our two major assumptions are: (1) questions following questions are assumed to be asking for clarification and not considered as having initiative, and (2) all statements following questions are assumed to answer these questions and are not considered as having initiative.

Although these assumptions are not always correct, defining a more precise annotation procedure is problematic. For some questions (*e.g.,* "what time is it?") annotators can agree what statements are answers and what statements give more information than requested. For other questions (*e.g.,* "what causes current to flow?") it is difficult to define a limited set of answers. Questions following questions are similarly difficult to deal with. Factors such as the content of the question, how many previous interruptions have occurred, and the response of the original speaker impact whether the question is perceived as taking initiative. Each factor is vague (*e.g.,* does the second interruption automatically take initiative?) and the interactions between them not well defined.

A second point is the strength of the results. The statistics presented here verify that the average amount of student initiative was higher in the didactic condition. If we refine our definition of initiative, it must change the data sufficiently such that the hypothesis that the didactic dialogues have more initiative is no longer valid, and the change in data will need to be drastic enough to verify the hypothesis that the Socratic dialogues have more initiative. Changing a few examples where our annotation rules failed will not be sufficient.

In section 2.2 we described initiative as a noisy measure; all student questions are treated the same no matter how deep or shallow. Thus, it seems likely that we simply need a finer measure of active learning. Such a measure will also allow us to revisit the question - "Are the Socratic dialogues really Socratic?". We showed that the Socratic dialogues were more interactive but if the tutor was asking mostly shallow questions then the dialogues should not be called Socratic.

The question taxonomy in (Graesser & Person, 1994) is one way to identify deep tutor and student questions. (Jordan & Siler, 2002) suggests going further and classifying student answers. Although a tutor may ask a shallow question the student may give more information than requested acting as if a deep question had been asked.

We also plan to study scaffolding based on the results in (Chi et al., 2001). Chi *et al.* performed a corpus collection similar to ours coming up with a set of didactic and Socratic dialogues. Learning gain results for the two groups were indistinguishable but Chi *et al.* found that tutors when in the didactic condition sometimes gave away post-test question answers in their explanations. Chi *et al.* cite a greater amount of scaffolding episodes and a greater amount of student reading as the reasons that the Socratic students were able to construct the answers that subjects in the didactic sessions were simply told.

Although initiative did not turn out to be related directly to learning gain, we saw that on average students had initiative for 10% of the time in Socratic dialogues and 21% of the time in didactic dialogues. The question is whether successful tutorial dialogue is necessarily mixed initiative. Currently most tutorial dialogue systems either never take initiative or never allow the student to have initiative. Our hypothesis is that mixed-initiative systems would be more successful. A dialogue system evaluation in the database retrieval domain (Chu-Carroll & Nickerson, 2000) found that a mixed-initiative dialogue manager resulted in higher user satisfaction and better task efficiency than a system-initiative dialogue manager.

This study does not say anything about user-initiative systems nor about the dialogue genre of tutoring. Thus, we plan to run experiments with our machine tutor comparing both user and system initiative dialogue management to mixed-initiative dialogue management. If our hypothesis is correct and mixed-initiative tutors perform better, then further work must be done to determine when tutors should take and give away initiative and what types of linguistic signals are used when giving or taking initiative. The latter phenomenon can be complicated as participants may enter into a negotiation subdialogue to settle who will take initiative.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, *4*(2), 167–207.

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. In *Educational Researcher*, Vol. 13, pp. 4–16.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, *22*(2), 249–254.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science*, *13*(2), 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science*, *18*(3), 439–477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from Human Tutoring. *Cognitive Science*, *25*, 471–533.

Chu-Carroll, J., & Brown, M. K. (1998). An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions. *User Modeling and User-Adapted Interaction*, *8*, 215–253. Special issue on Computational Models of Mixed Initiative Interaction.

Chu-Carroll, J., & Nickerson, J. S. (2000). Evaluating Automatic Dialogue Strategy Adaptation for a Spoken Dialogue System. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 202–209.

Doran, C., Aberdeen, J., Damianos, L., & Hirschman, L. (2001). Comparing Several Aspects of Human-Computer and Human-Human Dialogues. In *2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.*

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Freedman, R. (1996). *Interaction of Discourse Planning, Instructional Planning and Dialogue Management in an Interactive Tutoring System.* Ph.D. thesis, Northwestern University.

Graesser, A. C., & Person, N. K. (1994). Question Asking During Tutoring. *American Educational Research Journal*, *31*(1), 104–137.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology*, *9*, 495–522.

Grosz, B. J., & Sidner, C. L. (1986). Attention, Intensions, and the Structure of Discourse. *Computational Linguistics*, *12*(3), 175–204.

Jordan, P. W., & Di Eugenio, B. (1997). Control and Initiative in Collaborative Problem Solving Dialogues. In *AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interactions* Stanford, CA.

Jordan, P. W., & Siler, S. (2002). Control and Initiative in Computer-Mediated Human Tutoring Dialogues. In *Proceedings of the ITS'02 Workshop on Empirical Methods for Tutorial Dialogue Systems.*

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Beverly Hills, CA.

Linell, P., Gustavsson, L., & Juvonen, P. (1988). Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, *26*, 415–442.

Merrill, D. C., Reiser, B. J., & Landes, S. (1992a). Human tutoring: Pedagogical strategies and learning outcomes. Paper presented at the annual meeting of the American Educational Research Association.

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992b). Effective Tutoring Techniques: Comparison of Human tutors and intelligent tutoring systems. *Journal of the Learning Sciences*, *2*(3), 277–305.

Munro, A. (1994). Authoring interactive graphical models. In de Jong, T., Towne, D. M., & Spada, H. (Eds.), *The Use of Computer Models for Explication, Analysis and Experimental Learning.* Springer Verlag.

Pea, R. D., & Kurland, D. M. (1984). Logo programming and development of planning skills. Tech. rep. 16, Columbia University, New York, NY.

Pirolli, P., & Anderson, J. R. (1985). Recursive programming by children. *Canadian Journal of Psychology*, *39*, 240–272.

Rosé, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2000). A Comparative Evaluation of Socratic versus Didactic Tutoring. Tech. rep. LRDC-BEE-1, LRDC, University of Pittsburgh.

Shah, F. (1997). *Recognizing and Responding to Student Plans in an Intelligent Tutoring System: CIRCSIM-Tutor*. Ph.D. thesis, Illinois Institute of Technology.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an Analysis of Discourse: The English used by teachers and pupils.* Oxford University Press.

Strayer, S. E., & Heeman, P. A. (2001). Reconciling Initiative and Discourse Structure. In *2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.*

VanLehn, K., Siler, S., Murray, C., & Baggett, W. B. (1998). What Makes a Tutorial Event Effective?. In Gernsbacher, M. A., & Derry, S. (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, Madison, WI* Hillsdale, NJ. Erlbaum.

Walker, M. A., & Whittaker, S. (1990). Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 70–79.

Whittaker, S., & Stenton, P. (1988). Cues and Control in Expert-Client Dialogues. In *Proc. of the 26th Annual Meeting of the Association for Computational Linguistics (ACL-88)*, pp. 123–130.