The Utility of a Graphical Representation of Discourse Structure in Spoken Dialogue Systems

Mihai Rotaru
University of Pittsburgh
Pittsburgh, USA
mrotaru@cs.pitt.edu

Diane J. Litman
University of Pittsburgh
Pittsburgh, USA
litman@cs.pitt.edu

Abstract

In this paper we explore the utility of the Navigation Map (**NM**), a graphical representation of the discourse structure. We run a user study to investigate if users perceive the NM as helpful in a tutoring spoken dialogue system. From the users' perspective, our results show that the NM presence allows them to better identify and follow the tutoring plan and to better integrate the instruction. It was also easier for users to concentrate and to learn from the system if the NM was present. Our preliminary analysis on objective metrics further strengthens these findings.

1 Introduction

With recent advances in spoken dialogue system technologies, researchers have turned their attention to more complex domains (e.g. tutoring (Litman and Silliman, 2004; Pon-Barry et al., 2006), technical support (Acomb et al., 2007), medication assistance (Allen et al., 2006)). These domains bring forward new challenges and issues that can affect the usability of such systems: increased task complexity, user's lack of or limited task knowledge, and longer system turns.

In typical information access dialogue systems, the task is relatively simple: get the information from the user and return the query results with minimal complexity added by confirmation dialogues. Moreover, in most cases, users have knowledge about the task. However, in complex domains things are different. Take for example tutoring. A tutoring dialogue system has to discuss concepts, laws and relationships and to engage in complex subdialogues to correct user misconceptions. In addition, it is very likely that users of such systems are not familiar or are only partially familiar with the tutoring topic. The length of system turns can also be affected as these systems need to make explicit the connections between parts of the underlying task.

Thus, interacting with such systems can be characterized by an increased user cognitive load associated with listening to often lengthy system turns and the need to integrate the current information to the discussion overall (Oviatt et al., 2004).

We hypothesize that one way to reduce the user's cognitive load is to make explicit two pieces of information: the purpose of the current system turn, and how the system turn relates to the overall discussion. This information is implicitly encoded in the intentional structure of a discourse as proposed in the Grosz & Sidner theory of discourse (Grosz and Sidner, 1986).

Consequently, in this paper we propose using a graphical representation of the discourse structure as a way of improving the performance of complex-domain dialogue systems (note that graphical output is required). We call it the **Navigation Map** (**NM**). The NM is a dynamic representation of the discourse segment hierarchy and the discourse segment purpose information enriched with several features (Section 3). To make a parallel with geography, as the system "navigates" with the user through the domain, the NM offers a cartographic view of the discussion. While a somewhat similar graphical representation of the discourse structure has been explored in one previous study (Rich and Sidner, 1998), to our knowledge we are the first to test its benefits (see Section 6).

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 360–367,
Prague, Czech Republic, June 2007.

As a first step towards understanding the NM effects, here we focus on investigating whether users prefer a system with the NM over a system without the NM and, if yes, what are the NM usage patterns. We test this in a speech based computer tutor (Section 2). We run a within-subjects user study in which users interacted with the system both with and without the NM (Section 4).

Our analysis of the users' subjective evaluation of the system indicates that users prefer the version of the system with the NM over the version without the NM on several dimensions. The NM presence allows the users to better identify and follow the tutoring plan and to better integrate the instruction. It was also easier for users to concentrate and to learn from the system if the NM was present. Our preliminary analysis on objective metrics further strengthens these findings.

2 ITSPOKE

ITSPOKE (Litman and Silliman, 2004) is a state-of-the-art tutoring spoken dialogue system for conceptual physics. When interacting with ITSPOKE, users first type an essay answering a qualitative physics problem using a graphical user interface. ITSPOKE then engages the user in spoken dialogue (using head-mounted microphone input and speech output) to correct misconceptions and elicit more complete explanations, after which the user revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision.

All dialogues with ITSPOKE follow a question-answer format (i.e. system initiative): ITSPOKE asks a question, users answer and then the process is repeated. Deciding what question to ask, in what order and when to stop is hand-authored beforehand in a hierarchical structure. Internally, system questions are grouped in *question segments*.

In Figure 1, we show the transcript of a sample interaction with ITSPOKE. The system is discussing the problem listed in the upper right corner of the figure and it is currently asking the question Tutor_5. The left side of the figure shows the interaction transcript (not available to the user at runtime). The right side of the figure shows the NM which will be discussed in the next section.

Our system behaves as follows. First, based on the analysis of the user essay, it selects a question segment to correct misconceptions or to elicit more complete explanations. Next the system asks every question from this question segment. If the user answer is correct, the system simply moves on to the next question (e.g. $\text{Tutor}_2 \rightarrow \text{Tutor}_3$). For incorrect answers there are two alternatives. For simple questions, the system will give out the correct answer accompanied by a short explanation and move on to the next question (e.g. $\text{Tutor}_1 \rightarrow \text{Tutor}_2$). For complex questions (e.g. applying physics laws), ITSPOKE will engage into a *remediation subdialogue* that attempts to remediate user's lack of knowledge or skills (e.g. $\text{Tutor}_4 \rightarrow \text{Tutor}_5$). The remediation subdialogue for each complex question is specified in another question segment.

Our system exhibits some of the issues we linked in Section 1 with complex-domain systems. Dialogues with our system can be long and complex (e.g. the question segment hierarchical structure can reach level 6) and sometimes the system's turn can be quite long (e.g. Tutor_2). User's reduced knowledge of the task is also inherent in tutoring.

3 The Navigation Map (NM)

We use the Grosz & Sidner theory of discourse (Grosz and Sidner, 1986) to inform our NM design. According to this theory, each discourse has a discourse purpose/intention. Satisfying the main discourse purpose is achieved by satisfying several smaller purposes/intentions organized in a hierarchical structure. As a result, the discourse is segmented into discourse segments each with an associated discourse segment purpose/intention. This theory has inspired several generic dialogue managers for spoken dialogue systems (e.g. (Rich and Sidner, 1998)).

The NM requires that we have the discourse structure information at runtime. To do that, we manually annotate the system's internal representation of the tutoring task with discourse segment purpose and hierarchy information. Based on this annotation, we can easily construct the discourse structure at runtime. In this section we describe our annotation and the NM design choices we made. Figure 1 shows the state of the NM after turn Tutor_5 as the user sees it on the interface (NM line numbering is for exposition only). Note that Figure 1 is not a screenshot of the actual system interface. The NM is the only part from the actual system interface. Figure 2 shows the NM after turn Tutor_1.

We manually annotated each system question/explanation for its intention(s)/purpose(s). Note that some system turns have multiple inten-

TUTOR$_1$: To analyze this problem we will first describe the motion of the person and his keys while he is holding them. Then we will look at the motion of the person and his keys after he lets go of them.

Let's begin by looking at the motion of the man and his keys while he is holding them.

How does his velocity compare to that of his keys?

User answer$_1$: man's velocity is bigger [incorrect]

TUTOR$_2$: Well… The problem statement says that the man is holding his keys motionless in front of his face. That means they are both falling together and are not moving with respect to each other. Hence their velocities are the same.

Now let's see what happens when he releases them.

To do so we will first find the individual forces exerted on the man and his keys, then both net forces, both accelerations, and finally we will describe and compare the velocities.

So what are the forces exerted on the man after he releases his keys? Please, specify their directions.

User answer$_2$: gravity vertically down [correct]

TUTOR$_3$: Good. Because there is just one force, gravity's, it is trivial to determine the NET force. So what is the direction of the NET force on the person?

User answer$_3$: down [correct]

TUTOR$_4$: Ok. Considering the net force, does the person have an acceleration? If yes, specify its direction.

User answer$_4$: the man has no acceleration [incorrect]

TUTOR$_5$: Well… To analyze the man's acceleration we will use Newton's second law. What is the definition of this law?

***Problem**: Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.*

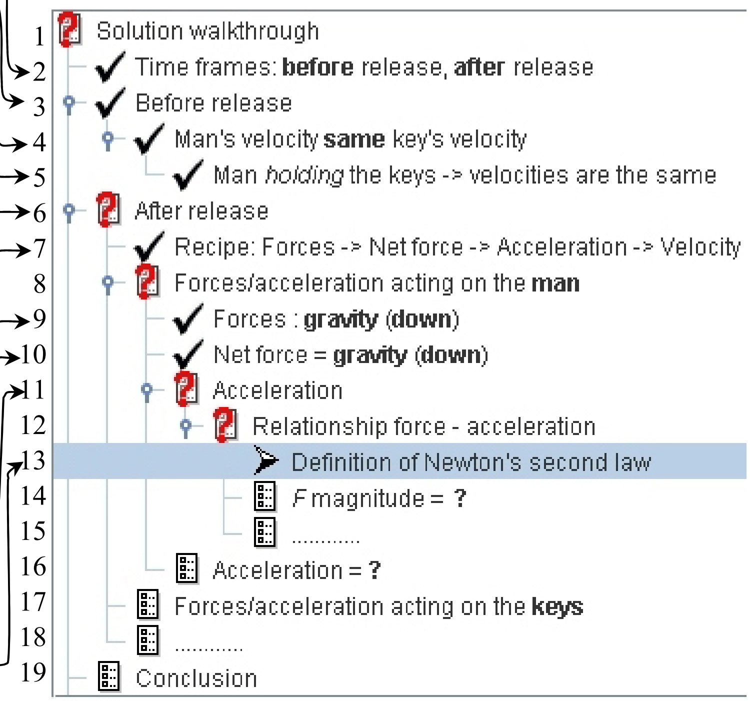

Figure 1. Transcript of a sample ITSPOKE speech interaction (left). The NM as the user sees it after turn Tutor$_5$

tions/purposes thus multiple discourse segments were created for them. For example, in Tutor$_1$ the system first identifies the time frames on which the analysis will be performed (Figure 1&2, NM$_2$). Next, the system indicates that it will discuss about the first time frame (Figure 1&2, NM$_3$) and then it asks the actual question (Figure 2, NM$_4$).

Thus, in addition to our manual annotation of the discourse segment purpose information, we manually organized all discourse segments from a question segment in a hierarchical structure that reflects the discourse structure.

At runtime, while discussing a question segment, the system has only to follow the annotated hierarchy, displaying and highlighting the discourse segment purposes associated with the uttered content. For example, while uttering Tutor$_1$, the NM will synchronously highlight NM$_2$, NM$_3$ and NM$_4$. Remediation question segments (e.g. NM$_{12}$) or explanations (e.g. NM$_5$) activated by incorrect answers are attached to the structure under the corresponding discourse segment.

3.1 NM Design Choices

In our graphical representation of the discourse structure, we used a left to right indented layout. In addition, we made several design choices to enrich the NM information content and usability.

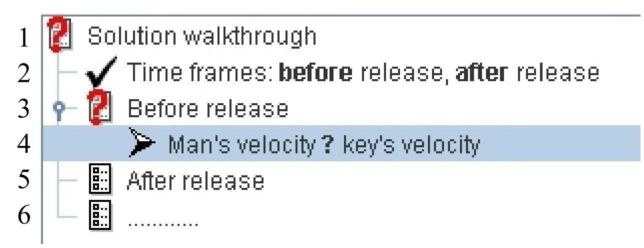

Figure 2. NM state after turn Tutor$_1$

Correct answers. In Figure 2 we show the state of the NM after uttering Tutor$_1$. The current discourse segment purpose (NM$_4$) indicates that the system is asking about the relationship between the two velocities. While we could have kept the same information after the system was done with this discourse segment, we thought that users will benefit from having the correct answer on the screen (recall NM$_4$ in Figure 1). Thus, the NM was enhanced to display the correct answer after the system is done with each question. We extracted the correct answer from the system specifications for each question and manually created a new version of the discourse segment purpose that includes this information.

Limited horizon. Since in our case the system drives the conversation (i.e. system initiative), we always know what questions would be discussed next. We hypothesized that by having access to this information, users will have a better idea of where instruction is heading, thus facilitating their understanding of the relevance of the current topic to the overall discussion. To prevent information overload, we only display the next discourse segment purpose at each level in the hierarchy (see Figure 1, NM_{14}, NM_{16}, NM_{17} and NM_{19}; Figure 2, NM_5); additional discourse segments at the same level are signaled through a dotted line. To avoid helping the students answer the current question in cases when the next discourse segment hints/describes the answer, each discourse segment has an additional purpose annotation that is displayed when the segment is part of the visible horizon.

Auto-collapse. To reduce the amount of information on the screen, discourse segments discussed in the past are automatically collapsed by the system. For example, in Figure 1, NM Line 3 is collapsed in the actual system and Lines 4 and 5 are hidden (shown in Figure1 to illustrate our discourse structure annotation.). The user can expand nodes as desired using the mouse.

Information highlight. Bold and italics font were used to highlight important information (what and when to highlight was manually annotated). For example, in Figure 1, NM_2 highlights the two time frames as they are key steps in approaching this problem. Correct answers are also highlighted.

We would like to reiterate that the goal of this study is to investigate if making certain types of discourse information explicitly available to the user provides any benefits. Thus, whether we have made the optimal design choices is of secondary importance. While, we believe that our annotation is relatively robust as the system questions follow a carefully designed tutoring plan, in the future we would like to investigate these issues.

4 User Study

We designed a user study focused primarily on user's perception of the NM presence/absence. We used a within-subject design where each user received instruction both with and without the NM.

Each user went through the same experimental procedure: 1) read a short document of background material, 2) took a pretest to measure initial physics knowledge, 3) worked through 2 problems with ITSPOKE 4) took a posttest similar to the pretest, 5) took a **NM survey**, and 6) went through a brief open-question interview with the experimenter.

In the 3rd step, the NM was enabled in *only one* problem. Note that in both problems, users did not have access to the system turn transcript. After each problem users filled in a **system questionnaire** in which they rated the system on various dimensions; these ratings were designed to cover dimensions the NM might affect (see Section 5.1). While the system questionnaire implicitly probed the NM utility, the NM survey from the 5th step explicitly asked the users whether the NM was useful and on what dimensions (see Section 5.1)

To account for the effect of the tutored problem on the user's questionnaire ratings, users were randomly assigned to one of two conditions. The users in the first condition (**F**) had the NM enabled in the first problem and disabled in the second problem, while users in the second condition (**S**) had the opposite. Thus, if the NM has any effect on the user's perception of the system, we should see a *decrease* in the questionnaire ratings from problem 1 to problem 2 for *F* users and an *increase* for *S* users.

Other factors can also influence our measurements. To reduce the effect of the text-to-speech component, we used a version of the system with human prerecorded prompts. We also had to account for the amount of instruction as in our system the top level question segment is tailored to what users write in the essay. Thus the essay analysis component was disabled; for all users, the system started with the same top level question segment which assumed no information in the essay. Note that the actual dialogue depends on the correctness of the user answers. After the dialogue, users were asked to revise their essay and then the system moved on to the next problem.

The collected corpus comes from 28 users (13 in *F* and 15 in *S*). The conditions were balanced for gender (*F*: 6 male, 7 female; *S*: 8 male, 7 female). There was no significant differences between the two conditions in terms of pretest ($p<0.63$); in both conditions users learned (significant difference between pretest and posttest, $p<0.01$).

5 Results

5.1 Subjective metrics

Our main resource for investigating the effect of the NM was the system questionnaires given after each problem. These questionnaires are identical and include 16 questions that probed user's perception of ITSPOKE on various dimensions. Users were asked to answer the questions on a scale from 1-5 (1 – Strongly Disagree, 2 – Disagree, 3 – Somewhat Agree, 4 – Agree, 5 – Strongly Agree). If indeed the NM has any effect we should observe differences between the ratings of the NM problem and the noNM problem (i.e. the NM is disabled).

Table 1 lists the 16 questions in the questionnaire order. The table shows for every question the average rating for all condition-problem combinations (e.g. column 5: condition *F* problem 1 with the NM enabled). For all questions except Q7 and Q11 a higher rating is better. For Q7 and Q11 (italicized in Table 1) a *lower* rating is better as they gauge negative factors (high level of concentration and task disorientation). They also served as a deterrent for negligence while rating.

To test if the NM presence has a significant effect, a repeated-measure ANOVA with between-subjects factors was applied. The within-subjects factor was the NM presence (**NMPres**) and the between-subjects factor was the condition (**Cond**)[1]. The significance of the effect of each factor and their combination (NMPres*Cond) is listed in the table with significant and trend effects highlighted in bold (see columns 2-4). Post-hoc t-tests between the NM and noNM ratings were run for each condition ("s"/"t"marks significant/trend differences).

Results for Q1-6

Questions Q1-6 were inspired by previous work on spoken dialogue system evaluation (e.g. (Walker et al., 2000)) and measure user's overall perception of the system. We find that the NM presence significantly improves user's perception of the system in terms of their ability to concentrate on the instruction (Q3), in terms of their inclination to reuse the system (Q6) and in terms of the system's matching of their expectations (Q4). There is a trend that it was easier for them to learn from the NM enabled version of the system (Q2).

Results for Q7-13

Q7-13 relate directly to our hypothesis that users benefit from access to the discourse structure information. These questions probe the user's perception of ITSPOKE during the dialogue. We find that for 6 out 7 questions the NM presence has a significant/trend effect (Table 1, column 2).

Structure. Users perceive the system as having a structured tutoring plan significantly[2] more in the NM problems (Q8). Moreover, it is significantly easier for them to follow this tutoring plan if the NM is present (Q11). These effects are very clear for *F* users where their ratings differ significantly between the first (NM) and the second problem (noNM). A difference in ratings is present for *S* users but it is not significant. As with most of the *S* users' ratings, we believe that the NM presentation order is responsible for the mostly non-significant differences. More specifically, assuming that the NM has a positive effect, the *S* users are asked to rate first the poorer version of the system (noNM) and then the better version (NM). In contrast, *F* users' task is easier as they already have a high reference point (NM) and it is easier for them to criticize the second problem (noNM). Other factors that can blur the effect of the NM are domain learning and user's adaptation to the system.

Integration. Q9 and Q10 look at how well users think they integrate the system questions in both a forward-looking fashion (Q9) and a backward looking fashion (Q10). Users think that it is significantly easier for them to integrate the current system question to what will be discussed in the future if the NM is present (Q9). Also, if the NM is present, it is easier for users to integrate the current question to the discussion so far (Q10, trend). For Q10, there is no difference for *F* users but a significant one for *S* users. We hypothesize that domain learning is involved here: *F* users learn better from the first problem (NM) and thus have less issues solving the second problem (noNM). In contrast, *S* users have more difficulties in the first problem (noNM), but the presence of the NM eases their task in the second problem.

Correctness. The correct answer NM feature is useful for users too. There is a trend that it is easier for users to know the correct answer if the NM is present (Q13). We hypothesize that speech recognition and language understanding errors are re-

[1] Since in this version of ANOVA the NM/noNM ratings come from two different problems based on the condition, we also run an ANOVA in which the within-subjects factor was the problem (Prob). In this case, the NM effect corresponds to an effect from Prob*Cond which is identical in significance with that of NMPres.

[2] We refer to the significance of the NMPres factor (Table 1, column 2). When discussing individual experimental conditions, we refer to the post-hoc t-tests.

Question	ANOVA			Average rating: F condition			S condition		
	NMPres	Cond	NMPres* Cond	P1 NM		P2 noNM	P2 NM		P1 noNM
Overall									
1. The tutor increased my understanding of the subject	0.518	0.898	0.862	4.0	>	3.9	4.0	>	3.9
2. It was easy to learn from the tutor	**0.100**	0.813	0.947	3.9	>	3.6	3.9	>	3.5
3. The tutor helped me to concentrate	**0.016**	0.156	0.854	3.5	>	3.0	3.9	$>^t$	3.4
4. The tutor worked the way I expected it to	**0.034**	0.886	0.157	3.5	>	3.4	3.9	$>^s$	3.1
5. I enjoyed working with the tutor	0.154	0.513	0.917	3.5	>	3.2	3.7	>	3.4
6. Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly	**0.004**	0.693	0.988	3.7	$>^s$	3.2	3.5	$>^s$	3.0
During the conversation with the tutor:									
7. ... a high level of concentration is required to follow the tutor	**0.004**	0.534	0.545	3.5	$<^s$	4.2	3.9	$<^t$	4.3
8. ... the tutor had a clear and structured agenda behind its explanations	**0.008**	0.340	0.104	4.4	$>^s$	3.6	4.3	>	4.1
9. ... it was easy to figure out where the tutor's instruction was leading me	**0.017**	0.472	0.593	4.0	$>^s$	3.4	4.1	>	3.7
10. ... when the tutor asked me a question I knew why it was asking me that question	**0.054**	0.191	**0.054**	3.5	~	3.5	4.3	$>^s$	3.5
11. ... it was easy to loose track of where I was in the interaction with the tutor	**0.012**	0.766	**0.048**	2.5	$<^s$	3.5	2.9	<	3.0
12. ... I knew whether my answer to the tutor's question was correct or incorrect	0.358	0.635	0.804	3.5	>	3.3	3.7	>	3.4
13. ... whenever I answered incorrectly, it was easy to know the correct answer after the tutor corrected me	**0.085**	**0.044**	0.817	3.8	>	3.5	4.3	>	3.9
At the end of the conversation with the tutor:									
14. ... it was easy to understand the tutor's main point	**0.071**	**0.056**	0.894	4.0	>	3.6	4.4	>	4.1
15. ... I knew what was wrong or missing from my essay	0.340	0.965	0.340	3.9	~	3.9	3.7	<	4.0
16. ... I knew how to modify my essay	0.791	0.478	0.327	4.1	>	3.9	3.7	<	3.8

Table 1. System questionnaire results

sponsible for the non-significant NM effect on the dimension captured by Q12.

Concentration. Users also think that the NM enabled version of the system requires less effort in terms of concentration (Q7). We believe that having the discourse segment purpose as visual input allows the users to concentrate more easily on what the system is uttering. In many of the open question interviews users stated that it was easier for them to listen to the system when they had the discourse segment purpose displayed on the screen.

Results for Q14-16

Questions Q14-16 were included to probe user's post tutoring perceptions. We find a trend that in the NM problems it was easier for users to understand the system's main point (Q14). However, in terms of identifying (Q15) and correcting (Q16) problems in their essay the results are inconclusive. We believe that this is due to the fact that the essay interpretation component was disabled in this experiment. As a result, the instruction did not match the initial essay quality. Nonetheless, in the open-question interviews, many users indicated using the NM as a reference while updating their essay.

In addition to the 16 questions, in the system questionnaire after the second problem users were asked to choose which version of the system they preferred the most (i.e. the first or the second problem version). 24 out 28 users (86%) preferred the NM enabled version. In the open-question interview, the 4 users that preferred the noNM version (2 in each condition) indicated that it was harder for them to concurrently concentrate on the audio and the visual input (divided attention problem) and/or that the NM was changing too fast.

To further strengthen our conclusions from the system questionnaire analysis, we would like to note that users were not asked to directly compare the two versions but they were asked to individually rate two versions which is a noisier process (e.g. users need to recall their previous ratings).

The NM survey

While the system questionnaires probed users' NM usage indirectly, in the second to last step in the experiments, users had to fill a NM survey which explicitly asked how the NM helped them, if at all. The answers were on the same 1 to 5 scale. We find that the majority of users (75%-86%) agreed or strongly agreed that the NM helped them follow the dialogue, learn more easily, concentrate and update the essay. These findings are on par with those from the system questionnaire analysis.

5.2 Objective metrics

Our analysis of the subjective user evaluations shows that users think that the NM is helpful. We would like to see if this perceived usefulness is reflected in any objective metrics of performance. Due to how our experiment was designed, the effect of the NM can be reliably measured only in the first problem as in the second problem the NM is toggled[3]; for the same reason, we can not use the pretest/posttest information.

Our preliminary investigation[4] found several dimensions on which the two conditions differed in the first problem (*F* users had NM, *S* users did not). We find that if the NM was present the interaction was shorter on average and users gave more correct answers; however these differences are not statistically significant (Table 2). In terms of speech recognition performance, we looked at two metrics: AsrMis and SemMis (ASR/Semantic Misrecognition). A user turn is labeled as AsrMis if the output of the speech recognition is different from the human transcript (i.e. a binary version of Word Error Rate). SemMis are AsrMis that change the correctness interpretation. We find that if the NM was present users had fewer AsrMis and fewer SemMis (trend for SemMis, $p<0.09$).

In addition, a χ^2 dependency analysis showed that the NM presence interacts significantly with both AsrMis ($p<0.02$) and SemMis ($p<0.001$), with fewer than expected AsrMis and SemMis in the NM condition. The fact that in the second problem the differences are much smaller (e.g. 2% for AsrMis) and that the NM-AsrMis and NM-SemMis interactions are not significant anymore, suggests that our observations can not be attributed to a difference in population with respect to system's ability to recognize their speech. We hypothesize that these differences are due to the NM text influencing users' lexical choice.

Metric	*F* (NM)	*S* (noNM)	p
# user turns	21.8 (5.3)	22.8 (6.5)	0.65
% correct turns	72% (18%)	67% (22%)	0.59
AsrMis	37% (27%)	46% (28%)	0.46
SemMis	5% (6%)	12% (14%)	0.09

Table 2. Average (standard deviation) for objective metrics in the first problem

6 Related work

Discourse structure has been successfully used in non-interactive settings (e.g. understanding specific lexical and prosodic phenomena (Hirschberg and Nakatani, 1996) , natural language generation (Hovy, 1993), essay scoring (Higgins et al., 2004) as well as in interactive settings (e.g. predictive/generative models of postural shifts (Cassell et al., 2001), generation/interpretation of anaphoric expressions (Allen et al., 2001), performance modeling (Rotaru and Litman, 2006)).

In this paper, we study the utility of the discourse structure on the user side of a dialogue system. One related study is that of (Rich and Sidner, 1998). Similar to the NM, they use the discourse structure information to display a segmented interaction history (**SIH**): an indented view of the interaction augmented with purpose information. This paper extends over their work in several areas. The most salient difference is that here we investigate the benefits of displaying the discourse structure information for the users. In contrast, (Rich and Sidner, 1998) never test the utility of the SIH. Their system uses a GUI-based interaction (no speech/text input, no speech output) while we look at a speech-based system. Also, their underlying task (air travel domain) is much simpler than our tutoring task. In addition, the SIH is not always available and users have to activate it manually.

Other visual improvements for dialogue-based computer tutors have been explored in the past (e.g. talking heads (Graesser et al., 2003)). However, implementing the NM in a new domain requires little expertise as previous work has shown

[3] Due to random assignment to conditions, before the first problem the *F* and *S* populations are similar (e.g. no difference in pretest); thus any differences in metrics can be attributed to the NM presence/absence. However, in the second problem, the two populations are not similar anymore as they have received different forms of instruction; thus any difference has to be attributed to the NM presence/absence in this problem as well as to the NM absence/presence in the previous problem.

[4] Due to logging issues, 2 *S* users are excluded from this analysis (13 *F* and 13 *S* users remaining). We run the subjective metric analysis from Section 5.1 on this subset and the results are similar.

that naïve users can reliably annotate the information needed for the NM (Passonneau and Litman, 1993). Our NM design choices should also have an equivalent in a new domain (e.g. displaying the recognized user answer can be the equivalent of the correct answers). Other NM usages can also be imagined: e.g. reducing the length of the system turns by removing text information that is implicitly represented in the NM.

7 Conclusions & Future work

In this paper we explore the utility of the Navigation Map, a graphical representation of the discourse structure. As our first step towards understanding the benefits of the NM, we ran a user study to investigate if users perceive the NM as useful. From the users' perspective, the NM presence allows them to better identify and follow the tutoring plan and to better integrate the instruction. It was also easier for users to concentrate and to learn from the system if the NM was present. Our preliminary analysis on objective metrics shows that users' preference for the NM version is reflected in more correct user answers and less speech recognition problems in the NM version.

These findings motivate future work in understanding the effects of the NM. We would like to continue our objective metrics analysis (e.g. see if users are better in the NM condition at updating their essay and at answering questions that require combining facts previously discussed). We also plan to run an additional user study with a between-subjects experimental design geared towards objective metrics. The experiment will have two conditions: NM present/absent for all problems. The conditions will then be compared in terms of various objective metrics. We would also like to know which information sources represented in the NM (e.g. discourse segment purpose, limited horizon, correct answers) has the biggest impact.

Acknowledgements

This work is supported by NSF Grants 0328431 and 0428472. We would like to thank Shimei Pan, Pamela Jordan and the ITSPOKE group.

References

K. Acomb, J. Bloom, K. Dayanidhi, P. Hunter, P. Krogh, E. Levin and R. Pieraccini. 2007. *Technical Support Dialog Systems: Issues, Problems, and Solutions.* In Proc. of Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies.

J. Allen, G. Ferguson, B. N., D. Byron, N. Chambers, M. Dzikovska, L. Galescu and M. Swift. 2006. Chester: Towards a Personal Medication Advisor. *Journal of Biomedical Informatics, 39*(5).

J. Allen, G. Ferguson and A. Stent. 2001. *An architecture for more realistic conversational systems.* In Proc. of Intelligent User Interfaces.

J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner and C. Rich. 2001. *Non-Verbal Cues for Discourse Structure.* In Proc. of ACL.

A. Graesser, K. Moreno, J. Marineau, A. Adcock, A. Olney and N. Person. 2003. *AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head?* In Proc. of Artificial Intelligence in Education (AIED).

B. Grosz and C. L. Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics, 12*(3).

D. Higgins, J. Burstein, D. Marcu and C. Gentile. 2004. *Evaluating Multiple Aspects of Coherence in Student Essays.* In Proc. of HLT-NAACL.

J. Hirschberg and C. Nakatani. 1996. *A prosodic analysis of discourse segments in direction-giving monologues.* In Proc. of ACL.

E. Hovy. 1993. Automated discourse generation using discourse structure relations. *Articial Intelligence, 63*(Special Issue on NLP).

D. Litman and S. Silliman. 2004. *ITSPOKE: An intelligent tutoring spoken dialogue system.* In Proc. of HLT/NAACL.

S. Oviatt, R. Coulston and R. Lunsford. 2004. *When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns.* In Proc. of International Conference on Multimodal Interfaces.

R. Passonneau and D. Litman. 1993. *Intention-based segmentation: Human reliability and correlation with linguistic cues.* In Proc. of ACL.

H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark and S. Peters. 2006. Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. *International Journal of Artificial Intelligence in Education, 16.*

C. Rich and C. L. Sidner. 1998. COLLAGEN: A Collaboration Manager for Software Interface Agents. *User Modeling and User-Adapted Interaction, 8*(3-4).

M. Rotaru and D. Litman. 2006. *Exploiting Discourse Structure for Spoken Dialogue Performance Analysis.* In Proc. of EMNLP.

M. Walker, D. Litman, C. Kamm and A. Abella. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*.